International Journal of
Computer & Software Engineering

Review Article Open Access

Big Data Tools-An Overview

Rabie A. Ramadan

Computer Engineering Department, Cairo University, Giza, Egypt

Abstract

With the increasing of data to be analyzed either in social media, industry applications, or even science, there is a need for nontraditional methods for data analysis. Big data is a way for nontraditional strategies and techniques to organize, store, and process huge data collected from large datasets. Large dataset, in this context, means too large data that cannot be handled, stored, or processed using traditional tools and techniques or one computer. Therefore, there is a challenge to come up with different analytical approaches to analyze massive scale heterogeneous data coming with high speed. Consequently, big data has some characteristics that makes it different from any other data which are Veracity, Volume, Variety, Velocity, and Value. Veracity means variety of resources while Variety means data from different sources. Big data Value characteristic is one of the ultimate challenge that could be complex enough to be stored, extracted, and processed. The Volume deals with the size of the data and required storage while Velocity is related to data streaming time and latency. Throughout this paper, we review the state-of-the-art of big data tools. For the benefits of researchers, industry and practitioners, we review a large number of tools either commercial or free tools. The article also shows some of the important characteristics of the collected tools to make it easy on organizations and scientists to select the best tool for the type of data to be handled.

Publication History:

Received: July 27, 2017
Accepted: December 27, 2017
Published: December 29, 2017

Keywords:

Relational databases, Data warehouses, Data storage

Introduction

We are in the era of big data that it is changing the face of computation and networking. Meaningful information is extracted from historical as well as run time data become valuable. Traditional data systems, such as relational databases and data ware houses, have been used by the businesses and organizations form the last 40 years. Traditional systems are used with primarily structured data with the following characteristics:

1. Fields are clearly organized in records and records are stored in tables. Each table has a name as well as fields with some relations among the fields.
2. A design to retrieve data from a disk and load it into memory. With large volume of data, it is inefficient to load such data in memory and inefficient to handle it with small programs. In fact, relational and warehouse databases usually is capable of reading 8k or 16k block sizes only.
3. For managing and accessing the data, Structured Query Language (SQL) is used with the structured data.
4. The traditional storage solutions are too expensive.
5. Relational databases and data warehouses are not designed for the new types of data in terms of scale, storage and processing.

However, still some people ignore the concept of big data doubting that they can benefit from it. Nevertheless, some facts that may lead to the revolution of the big data concepts cannot be ignored. Some of these facts are:

1. The generated data in the past two years are more than the data created in the entire history.
2. Due to the fast increase in the generated data, it is expected by the year of 2020 to reach 1.7 megabytes of new data to be generated every second for every human being on the planet.
3. The universe digital data will be grown from 4.4 zettabyets to around 44 Zettabytes, or 44 trillion gigabytes.
4. The number of likes per minutes for brands and organizations on Facebook reached 34,722 likes.
5. 100 terabytes of data uploaded daily to Facebook only. Poor data across businesses and the government costs the U.S. economy $3.1 trillion dollars a year.

Big data term is defined as a collection of various unstructured, structured, and randomized data generated by multiple sources with different formats with huge volume that cannot be handled with current traditional data handling techniques. Big data is characterized by five parameters, shown in Figure 1 as follow [10]:

Volume

The major challenge of big data is the storage, formatting, and analysis of huge volume of data. Big data is not concerned of Giga Bytes anymore, it is now challenge of petabytes and zettabyets. In fact, by the year of 2020, it is expected that the data volume will grow by 50 times.

Variety

The variety means that data is unstructured or semi-structured data. For instance, the data could be in the form of documents, graphs, images, emails, audio, text, video, logs, and many other formats.

Velocity

The term velocity usually refers to the data streaming where data could be collected from sensors, video streaming, and websites. The problem with the data stream is data consistency and completeness as well as time and latency. Analyzing such data is still a challenge.

*Corresponding Author: Prof. Rabie A. Ramadan, Computer Engineering Department, Cairo University, Giza, Egypt; E-mail: rabie@rabieramadan.org

Citation: Ramadan RA (2017) Big Data Tools – An Overview. Int J Comput Softw Eng 2: 125. doi: https://doi.org/10.15344/2456-4451/2017/125

Value

Data value means how beneficial the data to be analyzed. In other words, is it worth to dig in the data? Such question is another challenge because it is costly in terms of money and time to analyze the data.

Veracity

Veracity is the data quality which refers to the data noise and accuracy. This is another factor that may affect the data analysis approach as well as the required time and data dependency in taking decisions.

Along with these parameters, there are some other issues such as performance [33], networking [46], pre-processing techniques [16], and analysis techniques [44]. This paper surveys some of the important and most used tools for big data. The main purpose of the paper is to categories the big data tools to the benefit of researchers and the big data community.

The paper is organized as follows: the following section discusses some of the big data challenges, section 3 explores data storage and management tools, section 4 explains the main features of some of the big data platforms, Big data analysis tools is discussed in section 5, section6explicatessome of the big data visualization tools, finally the paper ends with a discussion and conclusion.

Big data challenges

Big data has many challenges that can be classified according to Sivarajah et al. [49] to data challenges, process challenges, and management challenges. Data challenges involves the big data characteristics (volume, verity, velocity, value, and veracity) in addition to visualization and variability. In terms of data processing the challenges are in data acquisition and warehousing, data mining and cleaning, data aggregation and integration, analysis and modeling, and data interpretation. In terms of data management, security, privacy, data governances, data sharing, cost expeditious, and data ownership.

Loshin in [50] specified five more challenges as follow:

Uncertainty of the Data Management Landscape

There are many tools that are competing in working on the big data analysis and each one of them is related to specific technical area. So, the challenge is to choose the suitable tool among the large number of tools that does not introduce any unknown or risks the data.

The Big Data Talent Gap

Big data is a new hot topic and it seems that there are only few experts in the field. Although many people claim that they are experts, due to the novelty of the big data approaches and applications, there are a gap in the experts. In fact, finding a person with practical experience in handling big data is a challenge.

Getting Data into the Big Data Platform

With the huge data in terms of scale and variety along with unprepared data practitioner makes it hard to get the data into the big data platform. In addition to the challenge of accessibility and integration in such case.

Synchronization across Data Sources

With huge data feeded into the big data platforms, you may realize that this data is coming from variety of sources at different times

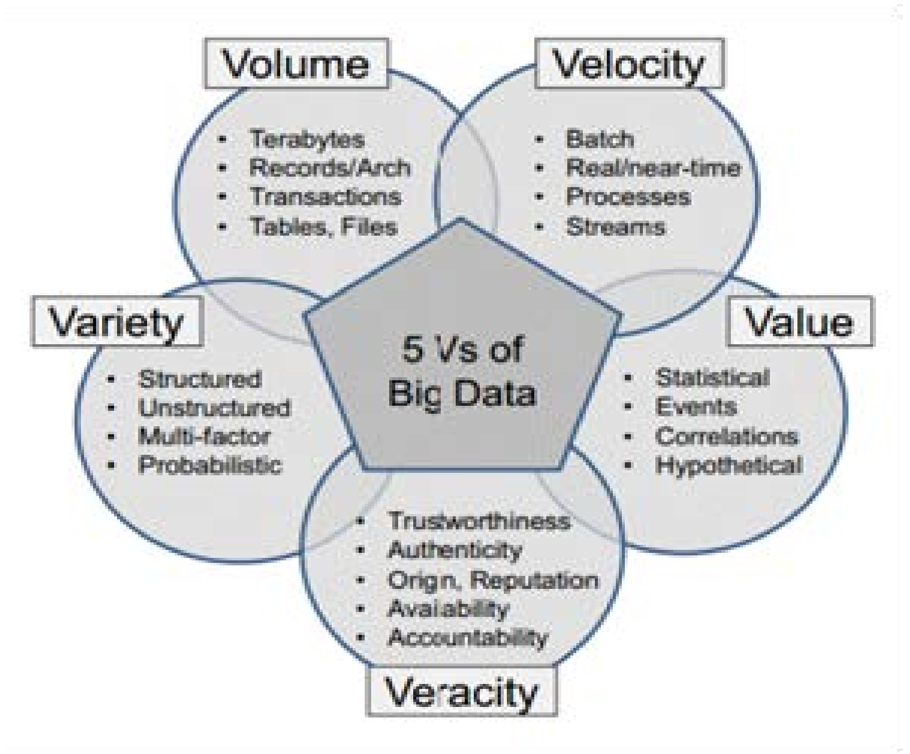


Figure 1. 5V's of big data [10]

with different rates. This may lead data to be out of synchronization with their sources. Without ensuring data synchronization, the data will be exposed to risk of analysis and inconsistency.

Getting Useful Information out of the Big Data Platform

Big data usually used by different applications for different purposes at different levels. The challenge here is to adequately provide these different applications with what they need to keep them doing what they supposed to do.

In the following sections, we will review some of the important big data tools and platforms classified into:

1. Data storage and management tools
2. Big Data analytics platforms
3. Big Data analysis tools
4. Big data visualization tools

Data storage and management tools

The problem with any application starts at the point of storing its huge collected data. Therefore, there are some solutions came to the light to solve the big data storage problem. The following are some of the most famous tools that are already used in some of the applications and Table 1 summarizes the characteristics of the data storage and management tools.

Cloudera [48]

Cloudera is an extension version of Hadoop with extra services. Cloudera has Enterprise Data Hub, analytical database, operational database, and data science and engineering products for treating different data. Enterprise Data Hub is an enterprise data hub that it is built on Hadoop framework and Apache Spark as the data repository. The hub is designed to deal with the big data issues in terms of volume, variety and velocity. It also supports some security techniques for sensitive data.

Apache Cassandra [4]

Cassandra is another distributed database management system over large structured database across multiple servers. Cassandra architecture is not like traditional architectures that uses master-slave or difficult-to-maintain shared architecture but it based on a masterless ring design where all of the nodes play the same role. Again, it is based on NoSQL quires where data is stored in no tabular or relational format. It is built on Dynamo-style replication model avoiding single point of failure; however, it adds a more powerful column family data model. Many of the companies like Apple, Comcast, eBay, Instagram, Spotify, Uber, and Netflix use Cassandra for their database core.

Chukwa [5,37]

Chukwa is another product that it is built on top of Hadoop. It is an open source distributed file system that inherits Hadoop Distributed File System (HDFS) and Map/Reduce framework features in terms of scalability and robustness. The core of Chukwa is a pipeline that take the data from it is generated to a place to be stored. It also involves toolkits for displaying, monitoring and analyzing data.

Apache HBase [2]

Apache HBase is an open-source platform that it is distributedfor non-relational database.HBase is designed after Google's Bigtable and Apache Phoenix project [https://phoenix.apache.org/] is used as its SQL layer. It also supports high table-update ratesand it maintains data consistency in reads and writes transactions. In addition, HBase features compression and inmemory operation.

MongoDB [30]

MongoDB is an open source documents database that is based on Value-Key pairs. The database uses NoSQL for query. It supports dynamic schema, full index, replication, high availability, and data interchange format called BSON. It also allows data to be distributed among different systems. The idea comes from the data explosion at Double Click web advertising company. MongoDBworks on Windows, Linux, OS X, and Solaris.

Tool	Type	Platform
Cloudera [48]	Hadoop distributed file system (HDFS)	Red Hat Enterprise Linux (RHEL), CentOS, Ubuntu, Debian
Apache Cassandra [4]	Database	Cross Platform
Chukwa[5, 37]	Hadoop distributed file system (HDFS)	Cross Platform
Apache HBase [2]	Hadoop distributed file system (HDFS)	Cross Platform
MongoDB [30]	Document-oriented database	Windows Vista and later, Linux, OS X 10.7 and later, Solaris, FreeBSD
Neo4j[11, 17, 31]	java – graph database	Cross-platform
CouchDB [6]	Erlang	Cross-platform
Terrastore [42]		
HibariDB[9]	Erlang - Key-value store	Cross-platform
Riak [38]	NoSQL database, cloud storage	Linux, BSD, macOS, Solaris
Hypertable[14]	associative array datastore / wide column store	Linux, Mac OS X
Blazegraph[3]	Graph	Ubuntu
Hive [25]	Data warehouse	Cross-platform
Infinispan [22]	data grid	Cross-platform

Table 1. Data storage and management characteristics

Neo4j [11,17,31]

It is open source community edition graph database. The main idea behind Neo4j is that it is built based on the concept of an edge, a node, or an attribute. Each node may have different attributes. Nodes and edges could have labels where they are used for search. Neo4j works on windows and Linux.

CouchDB [6]

CouchDB is another distributed database designed for big data. It is mainly designed for web that stores the data in JSON documents. CouchDB handles varying traffic including sudden spike in traffic. It is very scalable in terms of shrinking depending on the data needs. CouchDB works on Windows, Linux, OS X, and Android.

OrientDB [37]

OrientDB is a multi-model distributed NoSQL engine that works with different types of data including GeoSpatial, Document, Graph, Key-Value, and Reactive models. It is also scalable where it is based on layering architecture. It also involves security profiling based on users, roles, and search queries. OrientDB is written in java and works on O.S. independent.

Terrastore [42]

Terrastore as described at [42] is free tool based on Terracotta where it is build clustering methodology. It is also distributed, ubiquitous, elastic, scalable, consistent, and schemaless. It support different event processing, data partitioning, and range queries and it is OS Independent.

HibariDB [9]

Hibari is described as distributed, ordered key-value store. It is based on the concept of predictable latency especially in read and write ensuring data consistency and durability. Hibari is able to store PetaBytes of data distributing it across servers. In addition, it is based on ordered key-values. Moreover, it uses chain replication methodology for replicating the data across multiple servers. It is also O.S. Independent.

Riak [38]

Riak is a distributed NoSQL database. It is designed to work on for public, private, and hybrid clouds. So, it is recommended for IoT applications. The open source versions of Riak are Riak KV, Riak TS and Riak S2. It is based on NoSQL key-value data store in which it offers different features including: (1) high availability, (2) fault tolerance, (3) operational simplicity, and (4) scalability. Riak works on Linux and OS X.

Hypertable [14]

It is open source distributed database system that is completely implemented using C++. It is built on the top of Apache HDFS, GlusterFS or the CloudStore Kosmos File System (KFS). It is also based on NoSQL database offering efficient and fast data processing. It mainly works on Linux and OS X.

Blazegraph [3]

It is graph database that it is described as powerful and flexible representation of different types of data. It could be used in different types of applications including Internet of Things (IoT), Genomics / Biology, cyber security and national defense. It offers Blueprints and RDF/SPARQL APIs and supports for up to 50 Billion edges on a single machine. Blazegraph has a free version to use that it is platform independent.

Hive [25]

Hive is a data warehouse project that is built on the top of Hadoop for data summarization and analysis. It provides an SQL-like query language called HiveQL. It is developed by Facebook and other companies such as Netflix and Financial Industry Regulatory Authority (FINRA). Again it is O.S. independent.

Infinispan [22]

Infinispan is not quite big data database tool but it is a distributed in memory key/value data store with optional schema. It is written in java for language and O.S. independent. It also offers Web-based administration console for servers.

Tool	Description	Type of users	Availability
Hadoop [39]	It is an open source software that for data storage and distributed processing.	Expert	Open source and free
MaPR [26]	Platform consists of a set of data management tools and other relatedsoftware.	Expert	Converged Community Edition is free
IBM Big Data [21]	IBM platform offers many tools such as big SQL, Analytics for Apache Spark, BigIntegrate,Cloudant,Streams,Compose, and InfoSphere big match.	Expert	Commercial
KNIME Analytics [24]	It is an open solution for data-driven innovation, data mining, or predict new futures.	Average	The limited version is available as Open Source and Free
Microsoft Azure [28]	It is cloud based big data platform that it is for building, testing, deploying, and managing applications.	Average	Closed source for platform, Open source for client SDKs
Datameer [7]	It is a platform that allows easy loading and extracting data from multiple sources regardless the data formats and it is built on the top of Hadoop.	Average	Commercial
Amazon Web Service [1]	It provides a set of products for big data analysis such as compute, storage, databases, analytics, networking, mobile, developer tools, management tools, IoT, security and enterprise applications.	Average	Commercial

Table 2: Big Data Analytics Platforms

Big Data Analytics Platforms

Big data analytics platforms are enterprise platforms that are designed to provide functions and features for big data applications. It helps enterprises to discover unknown correlations, market trends, and useful information out of verity of big data. Table 2 summarizes the main characteristics of the most used platforms and the following is the description to them.

Hadoop [39]

Hadoop is an open source software that for data storage and distributed processing. The main idea behind Hadoop is store the data on distributed machines/clusters instead of keeping it on a single machine. Hadoop project started in 2005 and it is still one of the well-developed projects for big data storage. Hadoop framework is a little complicated where it consists of four modules which are Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce as can be seen in Figure 2.

Hadoop common contains libraries needed by other modules and Hadoop Distributed File System (HDFS) is a distributed file system on distributed machines. Hadoop YARN is responsible for managing

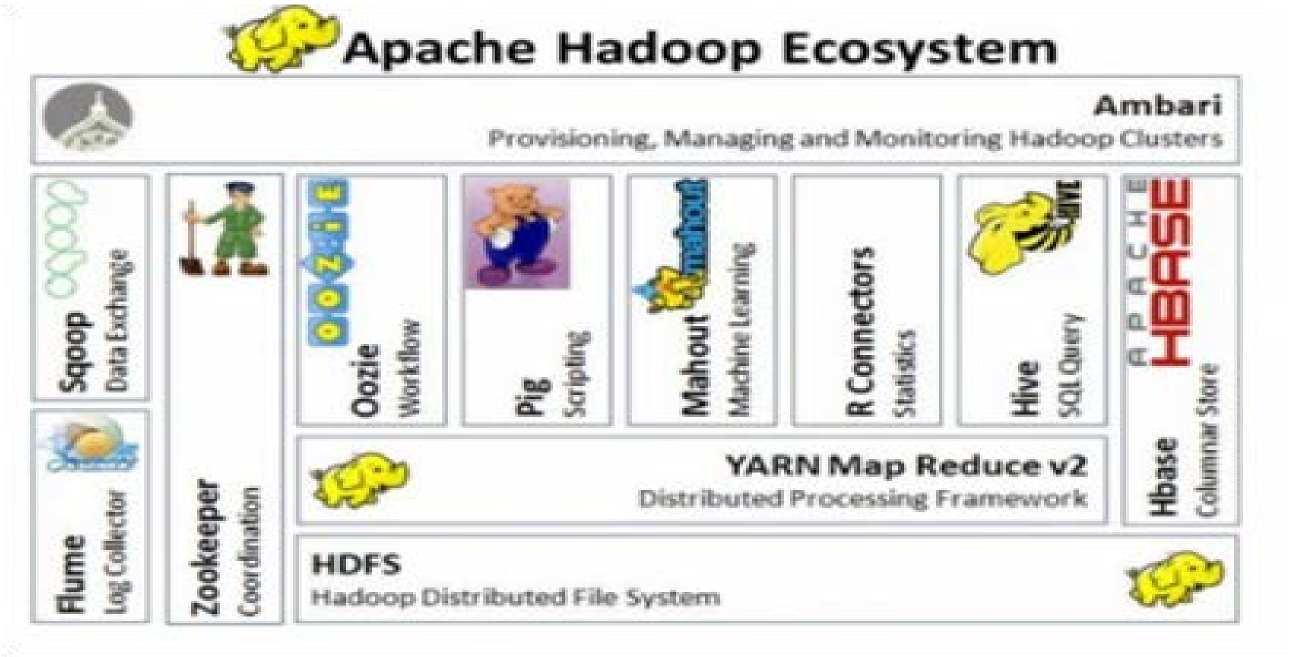


Figure 2: Hadoop Framework [39]

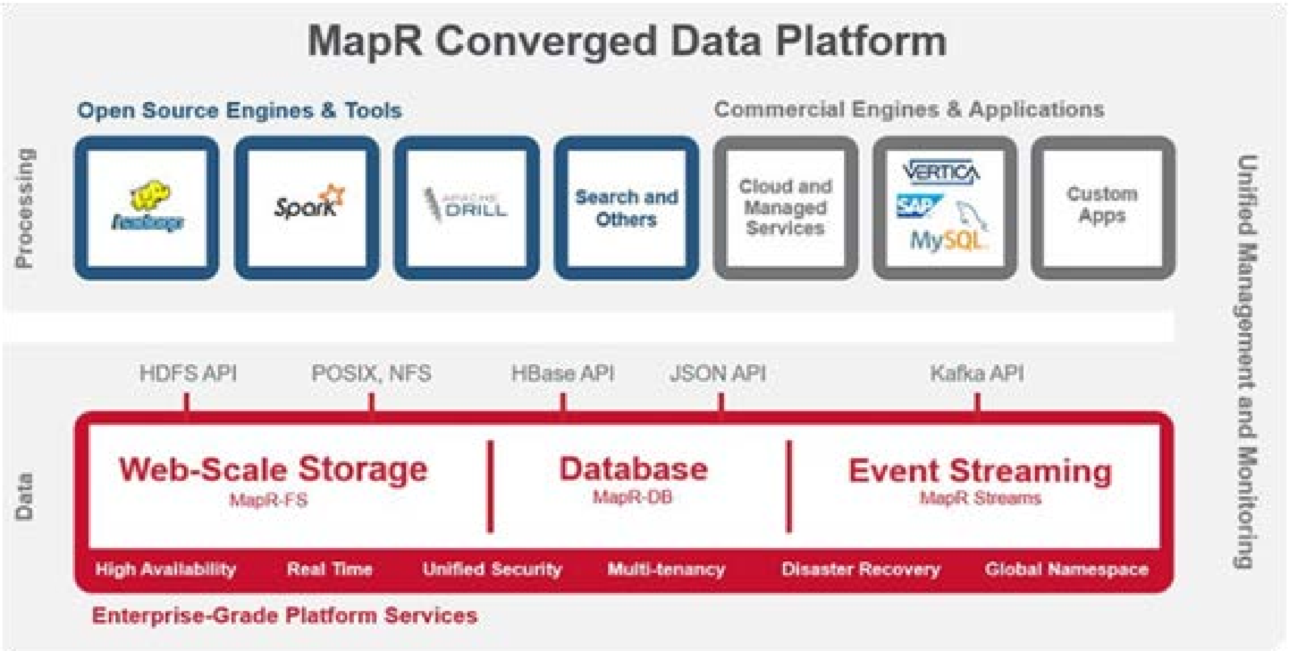


Figure 3. MapR architecture [26]

the system resources while Hadoop MapReduce is the programing model for the data processing which is in the Java language. One disadvantage of Hadoop is that it is hard to install and deal with.

MaPR [26]

The MapR is one of the open source platforms that integrates different tools such as Hadoop, Spark, and Apache Drill as shown in Figure 3. It handles real-time database, event streaming, and enterprise storage. In addition, it adds a security layer at the top of the data with reliable data processing. MapR has the following features:

1. MapR supports all Hadoop APIs and Hadoop data processing tools to access Hadoop data.
2. MapR provides true Network File System (NFS) capabilities.
3. MapR supports authentication services via Kerberos and/or LDAP.
4. Performance-optimized architecture for faster data processing and analytics
5. Architecture designed specifically for high availability across all cluster operations

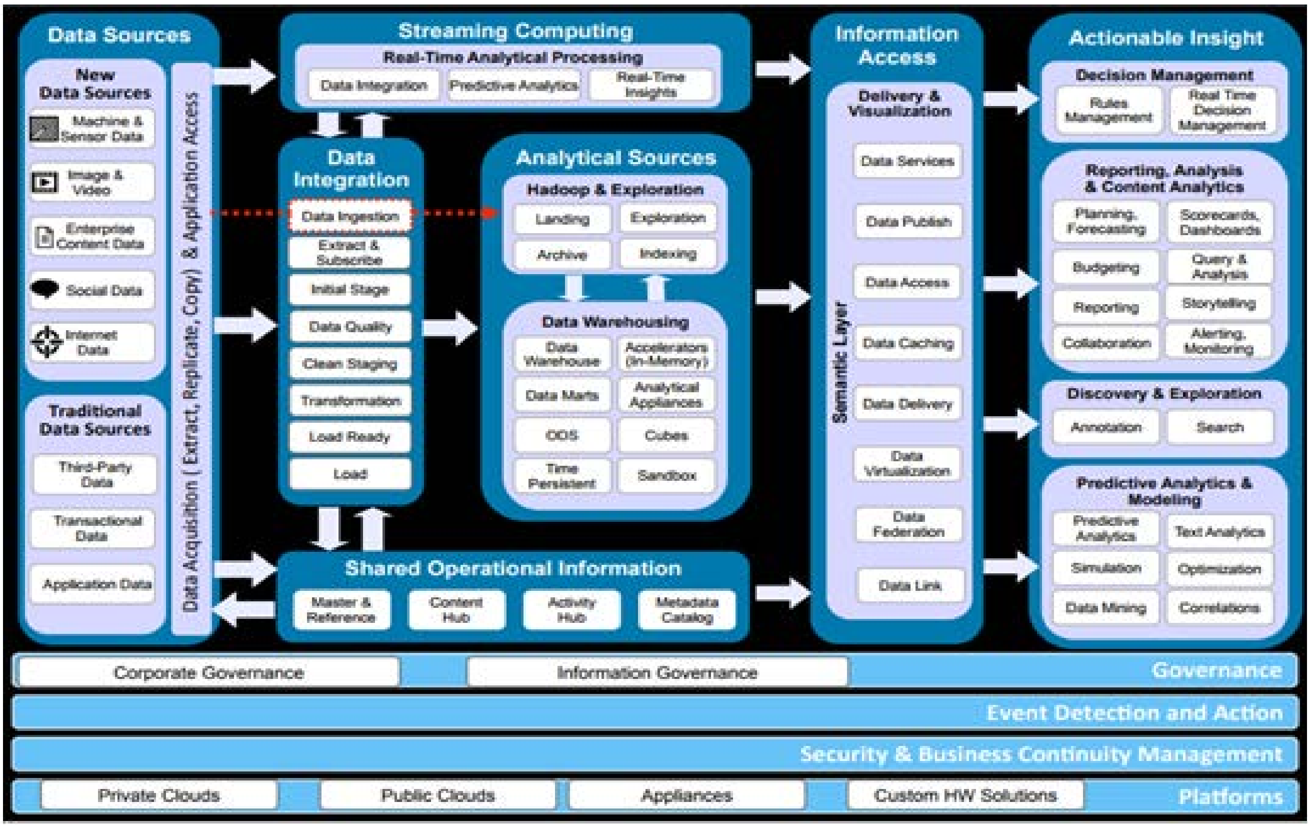


Figure 4. IBM Big Data and Analytics Reference Architecture Detailed Capabilities [20]

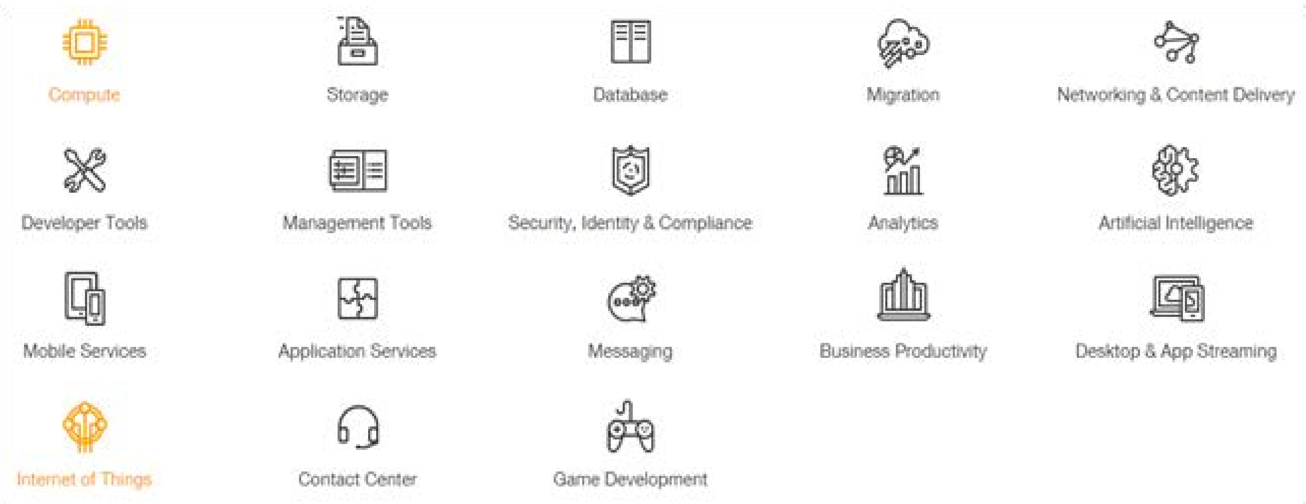


Figure 5. Amazon Big Data tools [50]

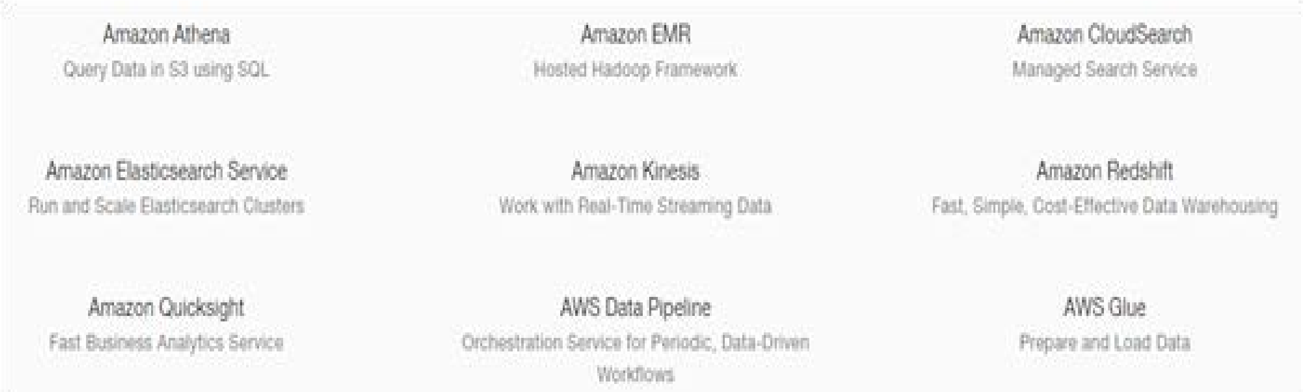
6. Automatic disaster recovery through mirroring to synchronize data across clusters
7. Direct Access NFS^{TM} for real-time data access to Hadoop data
8. Distributed metadata to support trillions of files in a single cluster
9. Comprehensive security controls to protect sensitive data
10. Consistent snapshots for accurate point-in-time recovery
11. MapR Heatmap^{TM} for instant cluster insights
12. MapR volumes for easier policy management around security, placement, retention, and quotas
13. Integrated NoSQL and event streaming for advanced real-time capabilities

IBM Big Data [21]

IBM offers many big data tools including IBM big SQL, Analytics for Apache Spark, BigIntegrate, Cloudant, Streams, Compose, and InfoSphere big match. It is designed to put forward enterprise Hadoop solutions, enable users to store big data, process data over multiple servers. At the same time, it makes the data accessible to all parties including IT users, business analysts, and data. Figure 4 is IBM Big Data and Analytics Reference Architecture –Detailed Capabilities.

Amazon Web Service [1]

Amazon provides another set of products for big data analysis including compute, storage, databases, analytics, networking, mobile, developer tools, management tools, IoT, security and enterprise


Amazon Athena
Query Data in S3 using SQL

Amazon EMR
Hosted Hadoop Framework

Amazon CloudSearch
Managed Search Service

Amazon Elasticsearch Service
Run and Scale Elasticsearch Clusters

Amazon Kinesis
Work with Real-Time Streaming Data

Amazon Redshift
Fast, Simple, Cost-Effective Data Warehousing

Amazon Quicksight
Fast Business Analytics Service

AWS Data Pipeline
Orchestration Service for Periodic, Data-Driven Workflows

AWS Glue
Prepare and Load Data

Figure 6. Amazon Analytics tool [1]

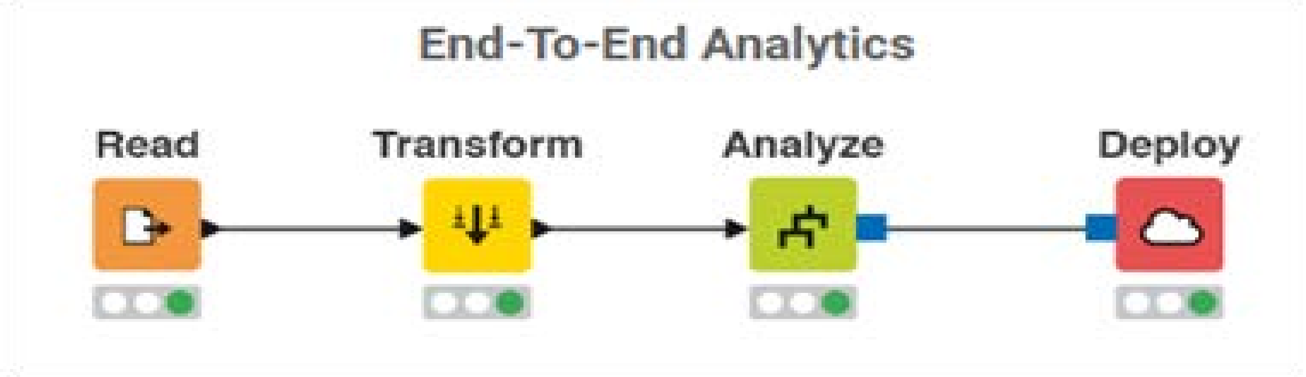


Figure 7. End to end module for KNIME [24]

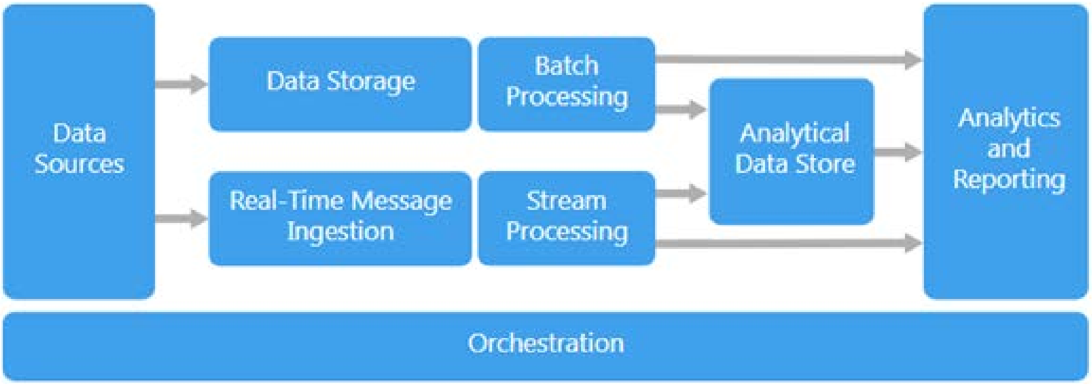


Figure 8. Microsoft Azurearchitecture [51]

applications, as shown in Figure 5. For instance, presented in Figure 6, analytics tool involves Amazon Athena which is query service allowing data analysis on Amazon S3 using standard SQL. AWS Elastic Beanstalk is an amazon service for web applications deployment and scaling using variety of programming languages and servers including Apache, Nginx, Passenger, and IIS.

KNIME Analytics [24]

KNIME analytics platform is an open source for data analytics that it is used to predict new futures and discover the potential hidden in the data. Figure 7 shows the end-to-end analytics proposed by KNIME. It contains more than 1000 modules and a range of advanced algorithms. Some of these modules are:

1. Connectors for all major file formats and databases
2. Support for a wealth of data types: XML, JSON, images, documents, and many more
3. Native and in-database data blending & transformation
4. Math & statistical functions
5. Advanced predictive and machine learning algorithms
6. Workflow control
7. Tool blending for Python, R, SQL, Java, Weka, and many more
8. Interactive data views & reporting

Microsoft Azure [28]

Microsoft Azure is cloud based big data platform that it is for building, testing, deploying, and managing applications. It provides different services including: (1) software as a service (SAAS), (2) platform as a service and (3) infrastructure as a service. Some of these services are HDInsight, Machine Learning, Stream Analytics, Azure Bot Service, Data Lake Analytics, Data Lake Store, and Data Catalog. HDInsight is a cloud Spark and Hadoop service while Machine Learning is a fully-managed cloud service that enables you to easily build, deploy, and share predictive analytics solutions. In addition, Stream Analytics is an on-demand real-time analytics service to power intelligent action while Azure Bot Service is an intelligent, serverless bot service that scales on demand. Data Lake AnalyticsData Lake Store, and Data Catalog are distributed analytics, hyperscale repository, and an enterprise-wide metadata catalog that makes data asset discovery straightforward, respectively. Big data solutions architecture provided by Microsoft Azure is shown in Figure 8.

Datameer[7]

Datameer is another platform that is built on the top of Hadoop as shown in Figure 9. Hadoop distributes the data and the computational load over multiple the network computers while the datameer tools allows easy loading and extracting data from multiple sources regardless the data formats. Datameer benefits from Hadoop

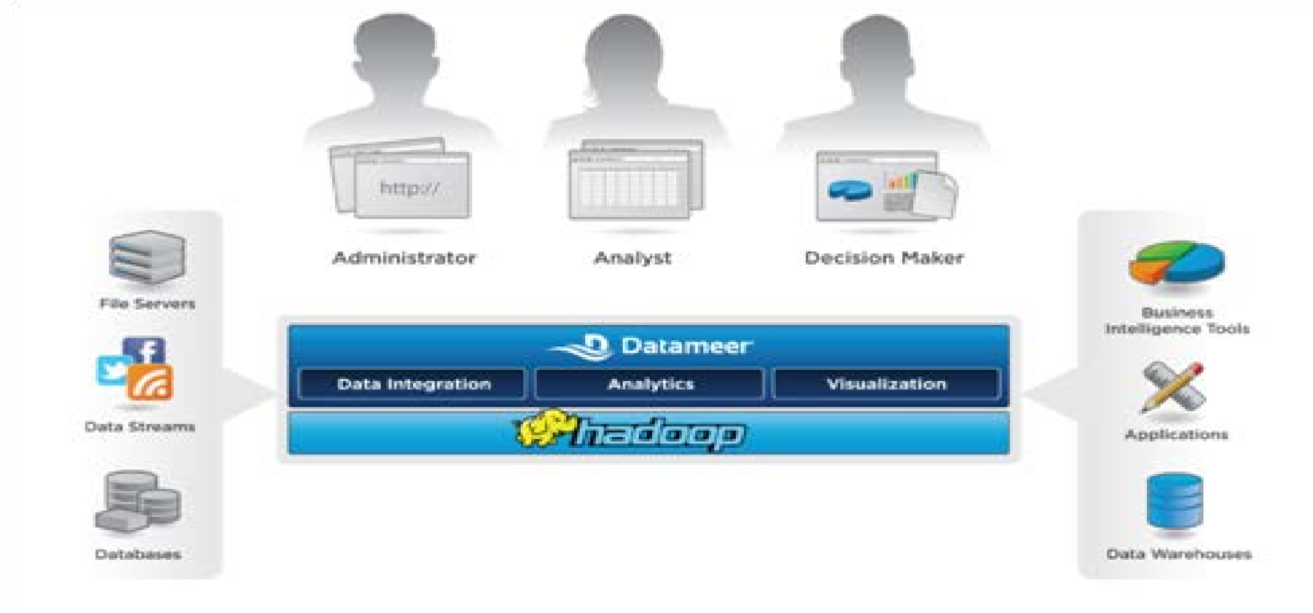


Figure 9. Datameer Architecture [52]

Tool	Description	Type of users	Availability
OpenRefine [43]	It is data cleaning tool.	Novice	Open source
R Project [36]	It is a statistical analysis and graphical language.	Average	Open source
GridGain [15]	It is in memory Computing Platform Built on Apache Ignite.	Average	Commercial
HPCC [19,29]	It is high performance data computing and parallel data processing for big data analysis.	Expert	Open source with some commercial modules
Apache Storm [18]	It is a distributed real-time computation system.	Expert	Open source

Table 3:Big Data Analysis Tools

in its data scaling and replication where Hadoop scales to 4000 servers and petabytes of data and the processes are fully parallelized inside Hadoop clusters. For data visualization there is built-in infographic widgets in addition to the interactive spreadsheet.

Big Data Analysis Tools

Data mining is all about searching of new patterns or unrecognized patterns. Data analysis is about breaking the big data down and tries to evaluate the impact of the discovered pattern over time. The following is the description to the most used big data analysis toolsand some of the characteristics of these tools are stated in Table 3.

OpenRefine (formerly Google Refine) [43]

OpenRefine is a data cleaning tool that it is easy to use. Its user interface is similar to Excel in which data can be imported or exported into different formats. It has built on a set of algorithms that it can match data together. It also allows direct connection to websites for uploading the resulted clean data. In addition, it is easy to learn tool but it need some time to review the cleaned data in case of large amount of data.

R Project [36]

R project is a statistical analysis and graphical language (linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering,). It is effective in terms of data handling and storage.Italsohasalargecollectionofanalysistools.Moreover,itisasimple programming language that it is very close to the traditional languages.

GridGain [15]

GridGain is in memory computing platform built on ApacheIgnite. It could be used as alternative to MapReduce since it is compatible with Hadoop Distributed File System. GridGrain is designed to work on real-time data on Windows, Linux, and OS X. It has three editions which are professional, enterprise, and ultimate editions. The difference between the professional and enterprise editions is that the professional does not support the following features:

1. Management & Monitoring GUI
2. Enterprise-Grade Security
3. Network Segmentation Protection
4. Rolling Production Updates
5. Data Center Replication
6. Cluster Snapshots

On the other hand, the enterprise edition doesn't support the cluster snapshot feature while the ultimate edition support it.

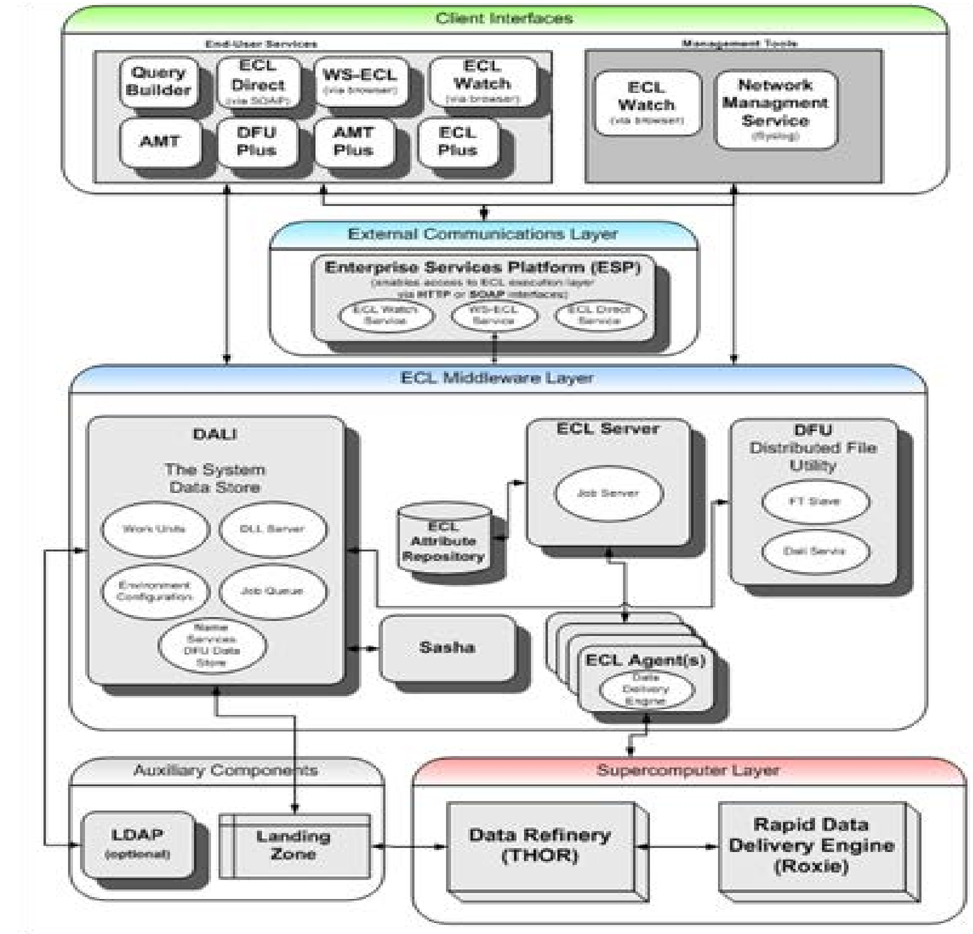


Figure 10. HPCCarchitecture [19]

HPCC [19, 29]

HPCC stands for High Performance Computing Cluster and is also knows as DAS (Data Analytics Supercomputer). HPCC is an open source data analysis tool developed by LexisNexis Risk Solutions providing high performance data computing and parallel data processing for big data analysis. HPCC architecture is given in Figure 10 where it depends on its analysis on patch parallel data processing and indexed data files (Roxie) format for data query. In case of parallel data processing HPCC uses one of two frameworks. The first framework is called a data refinery which is basically data cleaning and creation of keyed data and indexes for high performance computing while the second framework is designed for rapid data delivery engine.

Apache Storm [18]

Storm is another data analysis tool that it is owned by Twitter and described as Hadoop of realtime. Storm is developed to handle unbounded streams of data. Storm has no limitation in terms of development where any programming language can be used. At the same time, it is reliable where each tuple is processed at least once. Figure 11 shows the Storm streaming processing modules.

Big Data Visualization Tools

In this section, we introduce some of the most big data visualization tools including Google Fusion Tables, Tableau Public, Zoho Reports, Weave, Datawrapper, Jolicharts, Microsoft Power BI, Plotly, and Qlik Sense Desktop. The selection of these visualization tools are selected based on their functionalities and simplicity. Table 4 summarizes some of the features of these tools.

Google Fusion Tables [13]

Google fusion tables is one of the tools that it converts data into charts or maps as given in Figure 12. It allows users to upload files with different formats and presented in a chart or map forms including map,table, pie chart, bar graph, scatter plot, line chart and other types.

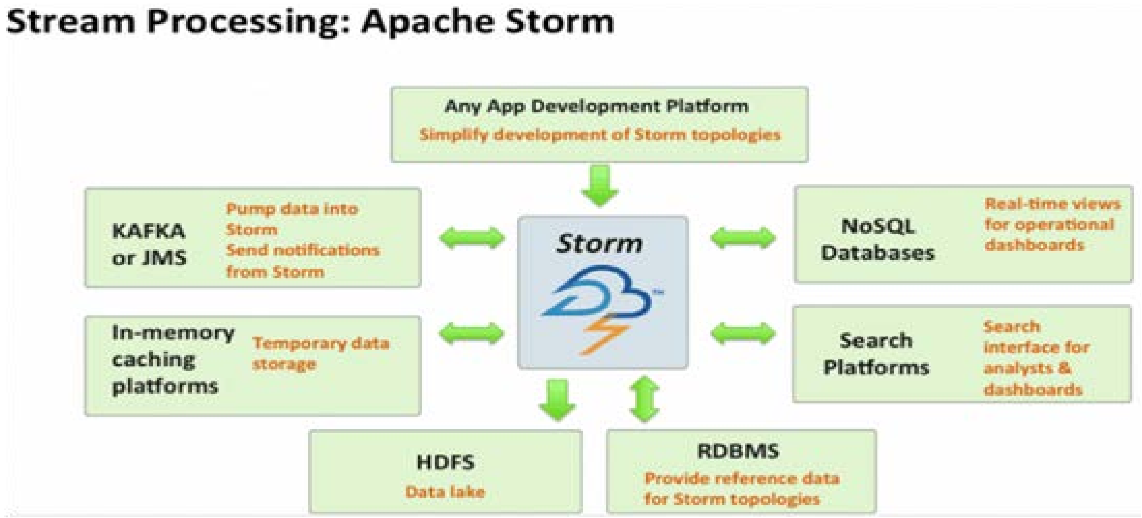
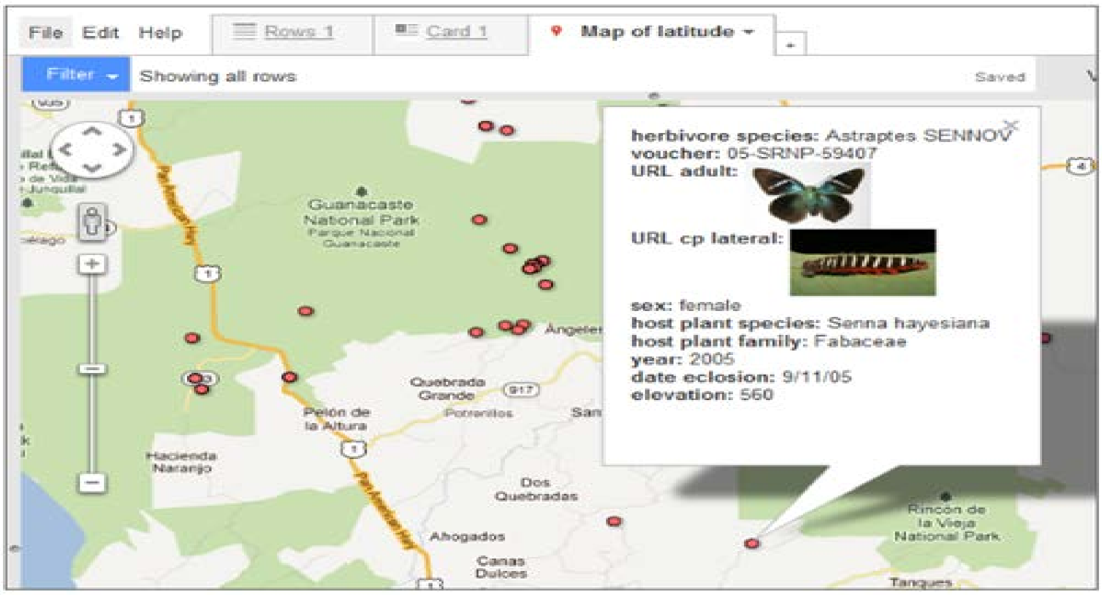

Stream Processing: Apache Storm
Any App Development Platform
Simplify development of Storm topologies
KAFKA or JMS
Pump data into Storm
Send notifications from Storm
Storm
NoSQL Databases
Real-time views for operational dashboards
In-memory caching platforms
Temporary data storage
Search Platforms
Search interface for analysts & dashboards
HDFS
Data lake
RDBMS
Provide reference data for Storm topologies

Figure 11. Storm streaming processing modules [18]



herbivore species: Astraptes SENNOV
voucher: 05-SRNP-59407
URL adult:
URL cp lateral:
sex: female
host plant species: Senna hayesiana
host plant family: Fabaceae
year: 2005
date eclosion: 9/11/05
elevation: 560

Figure 12. Google Fusion Table map sample [13].

It is easy to use and it utilizes geographic information system (GIS) functions to analyze data by geography. However, it is not a final version yet where google consider it as an experimental product.

Tableau Public [35]

Tableau public is another visualization tool that allows user visualize interactive data in suitable charts. It could be used by websites, students, writers, journalists, bloggers, and professors. It is easy tool to be used, so users can let the tool suggests the best visualization type. It also offers performing calculations on data and it allows almost 15 million row per workbook. However, the free version publish your work online even if it is not completed. In other words, the user has to complete his work in one session.

Zoho Reports [47]

Zoho reports is a web based tool for visualization that makes it easy on users to visualization different types of data file formats including databases. It converts the supported data files into charts or any formats suitable to the ones used by most spread sheets users. It also allow users to to create insightful reports and dashboards. However, the visualization tool still basic and limited especially in the free version where it supports only up to 100,000 row. Figure 13 shows a screenshot from the tool.

Weave [45]

Weave is an open source general-purpose visualization tool that allows users to create different visualizations including a bar chart, scatter plot and map. In addition, it handles wide variety of relational and no-relational data sources including SHP/DBF, SQL, CSV, GeoJSON, and CKAN. It is also very interactive tool that shows the runtime changes in the charts as well as it is a cross-platform tool. However, the tool requires experienced user to install it and requires flash enabled browser. Figure 14 shows a sample of Weave output.

Tool	Platform	API found	Data Storage	Type of data
Google Fusion Tables [13]	Web-Based	Yes	External Server	Chart, map, network graph, orcustom layout .
Tableau Public [35]	Windows, OS X	No	Public external server	relational databases, OLAP cubes, cloud databases, and spreadsheets
Zoho Reports [47]	Web-Based	Yes	External server	spreadsheets , flat files, stream data from online storage services, relational databases, and NoSQL databases
Weave [45]	Web-Based , Linux	Yes	Local , external server	CSV, relational databases, and geometric shape files.
Datawrapper [8]	Web-Based	Yes	Local , external server	CSV fileand Excel .
Jolicharts [23]	Web-Based	Yes	External server	Excel, Google Spreadsheets, Databases , and works for real-time data
Microsoft Power BI [27]	Web-Based, Windows	Yes	Local , External Server	Many data sources , big data, streaming data
Plotly [34]	Web-Based , scripting	Yes	Local , External Server	tabular data files (.xls, .xlsx, .csv, .tsv), Spectacle Presentation files (.spc, .json), and Jupyter Notebook files (.json, .ipynb)
Qlik Sense Desktop [40]	Windows, OS X	Yes	Local	Databases, CSV, Text files from remote files and web files
ChartBlocks [57]	Web-Based	No	External Server	Spreadsheets, databases and live datafeeds

Table 4. Summary of the visualization tools

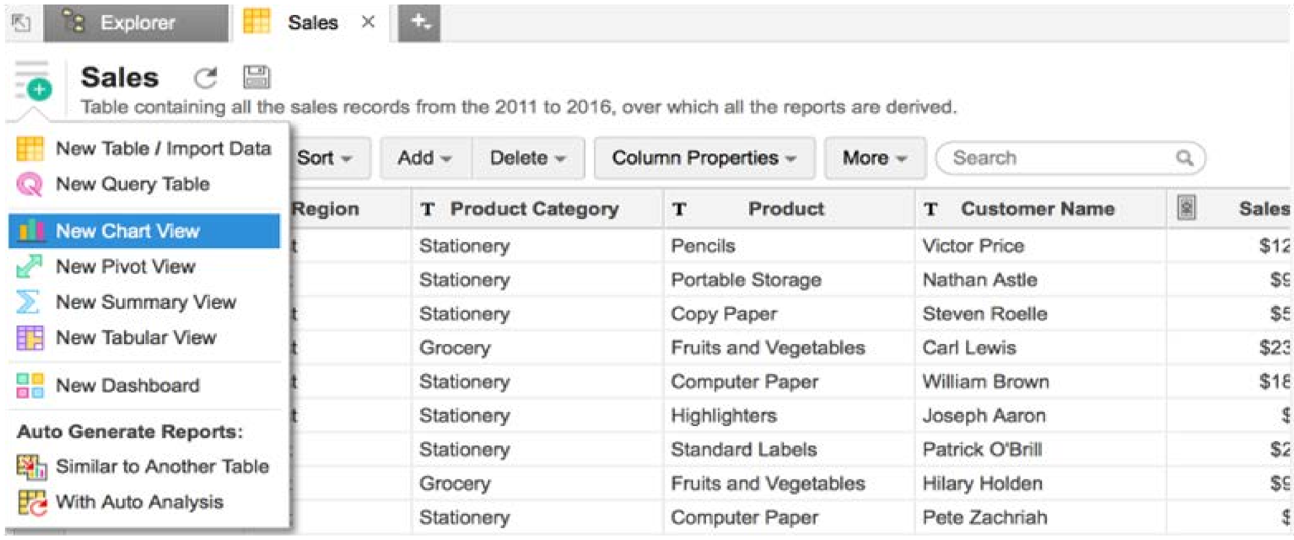


Table containing all the sales records from the 2011 to 2016, over which all the reports are derived.

Region	T Product Category	T Product	T Customer Name	Sales
t	Stationery	Pencils	Victor Price	$12
	Stationery	Portable Storage	Nathan Astle	$9
t	Stationery	Copy Paper	Steven Roelle	$5
t	Grocery	Fruits and Vegetables	Carl Lewis	$23
t	Stationery	Computer Paper	William Brown	$18
t	Stationery	Highlighters	Joseph Aaron	$
	Stationery	Standard Labels	Patrick O'Brill	$2
	Grocery	Fruits and Vegetables	Hilary Holden	$9
	Stationery	Computer Paper	Pete Zachriah	$

Figure 13. Zoho Reportssample [47]

Datawrapper [8]

Datawrapper is an open source tool that allows users to create simple and embedded charts as given in Figure 15. It is originally developed for journalist but it could be used by others as well. The charts are hosted on Amazon Web Services and users will be able to download the data through a provided link. Datawrapper offers control over colors and data grouping with no previous experience; however, it is missing the flexibility in customization. The free Datawrappersupports up to 10,000 chart views; however, the commercial version is unlimited.

Jolicharts [23]

Jolicharts is web based service that converts spreadsheets, JSON or databases into maps, graphs, or charts. Multiple charts on a dashboard can be arranged and it allows to add filter boxes multiple visualizations can be handled at once. Jolicharts allows only 50MB data storage for free; however, the commercial version allows data storage up to 250GB. In addition, the documentation needs enhancement.A sample chart generated by Jolicharts is presented in Figure 16.

Microsoft Power BI [27]

Power BI allows user to simply create interactive visualizations, dashboards and reports out of large data.It uses a simple natural language (R language) to query the data on dashboards. It is flexible to use multiple data sources including Salesforce, MailChimp, Google Analytics, and GitHub. Some of these sources are presented in Figure 17. In addition, Power BI has different versions for different platforms such as desktop, Mobile, web, and embedded. The desktop version is free while other versions are commercial.

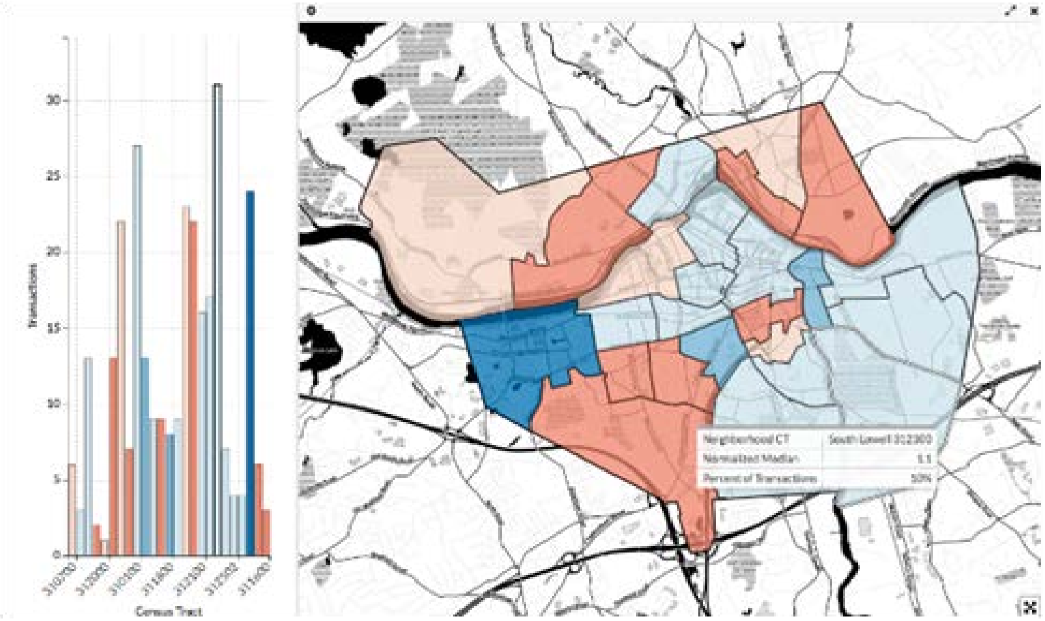


Figure 14. Weave project output sample [45].

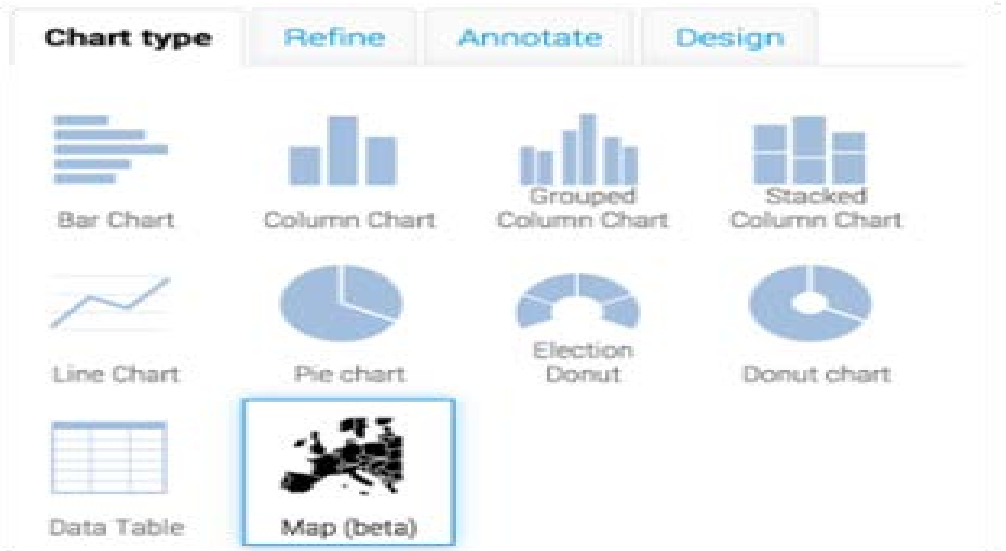


Figure 15. Datawrapper service sample [8].

Plotly is another visualization tool that allows creating and sharing data visualizations. In addition, it includes statistical analysis tools. Moreover, it offers separate API (R, Python, JavaScript and MATLAB APIs) for users to add their own custom graph functions. Again it accepts several data formats including CSV, Excel, and Access, spreadsheets from Google Drive, TSV, and MATLAB. The community version of Plotly involves only unlimited public files, PNG and JPEG data export, 250 API calls per day, and the connection to 7 data sources. However, the commercial version could support Unlimited Private Charts, Dashboards and Slide Decks, SVG, EPS, HTML, PDF, PNG and JPEG data export, 10,000 API Calls per day, and Connect to 18 Data Sources. Figure 18 shows a sample from Plotly charts.

Qlik Sense Desktop [40]

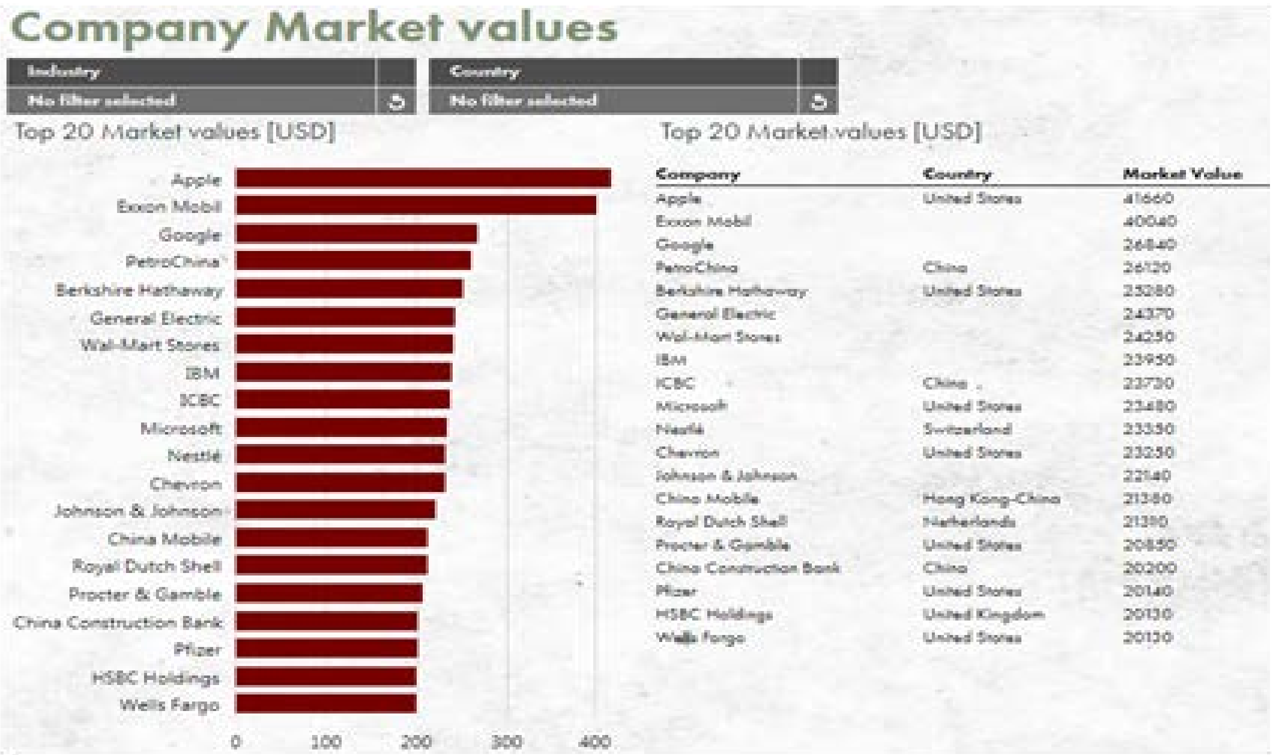
Qlik Sense Desktop is another visualization service that is offered as cloud based or desktop based. The service allows drag and drop to create interactive data visualization without complex SQL queries, scripting, or joins. It has an APIs for developers to extend and embed Qlik Sense in their applications. Figure 19 shows a sample from Qlik Sense dashboard.

ChartBlocks [57]:

ChartBlocks is an easy tool for generating different types of charts including Bar, Line, Pie, and Scatter. ChartBlocks provides an Application Programming Interface (API) library for creating charts and embedding them into web pages. The library allows creating stream live data charts as well as pulling raw data. ChartBlocks tool comes with four options including basic, personal, professional, and elite. The only free version is the basic which allows up to 30 active charts.

Discussion and Conclusion

Throughout this paper, we reviewed the most used big data techniques. These techniques are classified into four classes which



Company Market values

Top 20 Market values [USD]

Company	Country	Market Value
Apple	United States	41660
Exxon Mobil		40040
Google		26840
PetroChina	China	26120
Berkshire Hathaway	United States	25280
General Electric		24370
Wal-Mart Stores		24250
IBM		23950
ICBC	China	23730
Microsoft	United States	23480
Nestlé	Switzerland	23330
Chevron	United States	23250
Johnson & Johnson		22140
China Mobile	Hong Kong-China	21380
Royal Dutch Shell	Netherlands	21310
Procter & Gamble	United States	20850
China Construction Bank	China	20200
Pfizer	United States	20140
HSBC Holdings	United Kingdom	20130
Wells Fargo	United States	20130

Figure 16. Jolicharts sample chart [23].

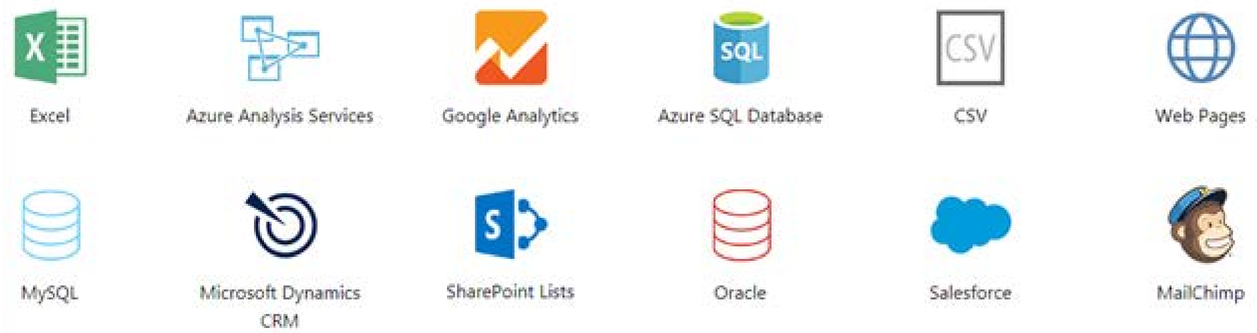
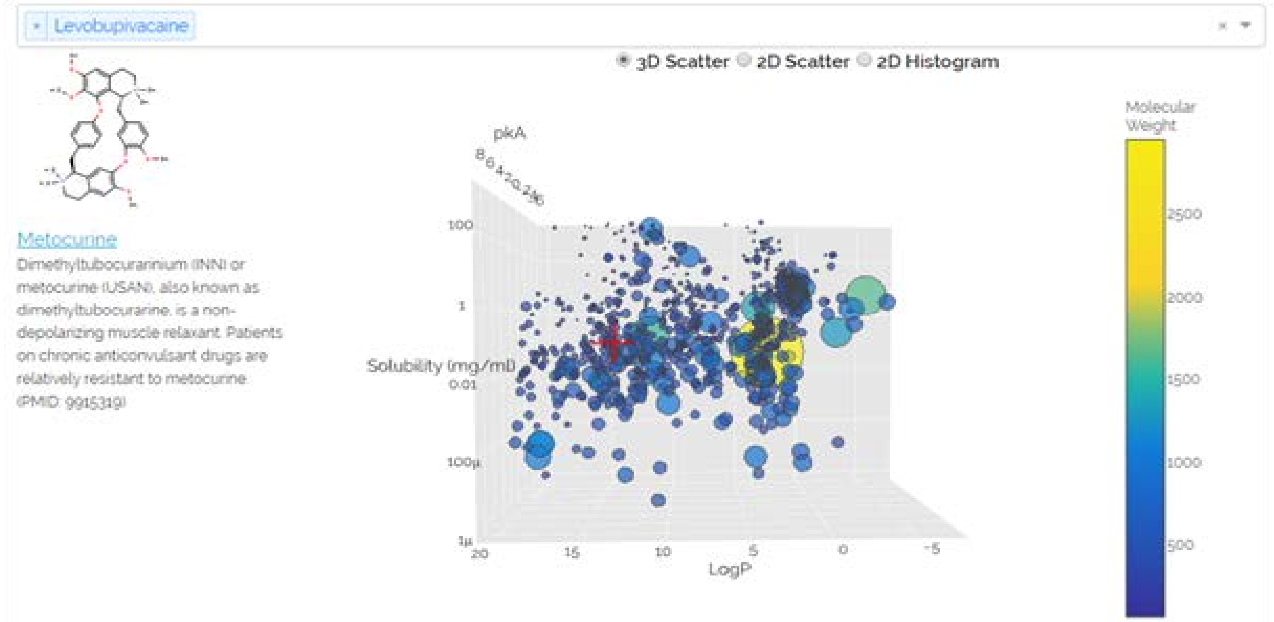


Figure 17. Microsoft Power BI data sources [27].

are storage and management, platforms, data analysis, and data visualization. As can be seen, these tools are gaining great attention from academia, businesses, and industry. They also cover a wide range of applications and they are very successful in many of these applications. However, our vision to the future could be summarized as follows [53-56]:

1. The learning curve of the current tools and platforms are high and it needs experienced users to deal with.
2. The available big data platforms are not generic enough to be used in different applications. In addition, the free versions are very limited. At the same time, the commercial versions are expensive and does not fit mid-size organizations as well as researchers.
3. The cost of the commercial tools limits Universities and Masters and Ph.D. students' contributions.
4. Although there are some tools are considered as "cross platform", the actual implementation and usage to these tools reveals that there are not straight forward, a large number of problems appear when using different platforms.
5. Although some tools are using some of the Computational Intelligence (CI) techniques in their cores, still there is a lack of implementation to a wide effective range of algorithms and techniques.
6. Up to our knowledge, most of the big data languages are new and not tested in real applications.

In our future work, we will investigate deeply into data storage and management tools. In addition, we will report our practical experience using each of these tools



Metocurine

Dimethyltubocurarinium (INN) or metocurine (USAN), also known as dimethyltubocurarine, is a non-depolarizing muscle relaxant. Patients on chronic anticonvulsant drugs are relatively resistant to metocurine (PMID: 9915319)

Figure 18. Sample from Plotly charts [34].

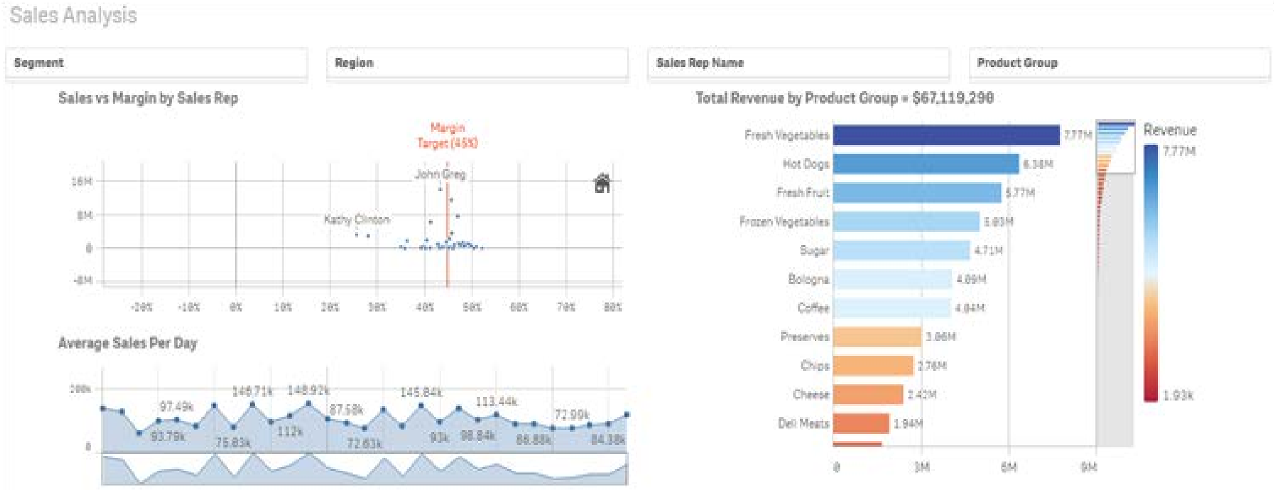


Figure 19. A sample from Qlik Sense dashboard [40].

Citation: Ramadan RA (2017) Big Data Tools – An Overview. Int J Comput Softw Eng 2: 125. doi: https://doi.org/10.15344/2456-4451/2017/125

Conflict of Intreast

No authors have a conflict of interest or any financial tie to disclose.

References

1. Amazon Web Services Inc (2017). Amazon Web Services.
2. Apache HBase (2017).
3. Blazegraph (2017) Blazegraph.
4. Cassandra (2016).
5. Chukwa (2016).
6. CouchDB (2017) CouchDB.
7. Datameer. (2017). Datameer.
8. Datawrapper (2017) Datawrapper.
9. DB H (2015) Hibari DB.
10. Demchenko Y, de Laat C, Membrey P (2014) Defining architecture components of the Big Data Ecosystem. In International Conference on Collaboration Technologies and Systems (CTS) pp. 104-112.
11. Eifrem E (2011) Graph Databases, Licensing and MySQL.
12. Foundation, T. A. S. (2016). cassandra.
13. Google.com. (2017). Google Fusion Tables.
14. Googlestack (2012) Taking Hypertable Out For A Spin.
15. GridGain (2017) GridGain.
16. Hariharakrishnan J, Mohanavalli S, Srividya, Kumar KBS. (2017) Survey of Pre-processing Techniques for Mining Big Data. International Conference on Computer, Communication and Signal Processing (ICCCSP) pp 1-5.
17. Hoff T (2009) Neo4j - a Graph Database that Kicks Buttox.
18. Hortonworks (2017) Storm.
19. HPCC (2017) HPCC Systems.
20. IBM (2014) IBM Big Data & Analytics Reference Architecture V1.
21. IBM. (2017). IBM Big Data.
22. Infinispan (2017) Infinispan.
23. Jolicharts (2017) Jolicharts.
24. KNIME (2017) KNIME Analytics Platform.
25. Leverenz L (2017) Hive.
26. MaPR (2017) MaPR.
27. Micorsoft BI (2017) Microsoft Power BI.
28. Microsoft. (2017). Microsoft Azure.
29. Middleton AM (2011). White Paper HPCC Systems : Introduction to HPCC (High-Performance Computing Cluster).\
30. MongoDB (2017)
31. Neo4j (2017) neo4j.
32. OrientDB (2017) OrientDB.
33. Pattanshetti T, Attar V (2017) Survey of performance modeling of big data applications. In 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence.
34. Plot.ly (2017) Plotly.
35. Public T (2017) Tableau Public.
36. R (2017) The R Project for Statistical Computing
37. Rabkin A, Katz R (2010) Chukwa: a system for reliable large-scale log collection. In LISA'10 Proceedings of the 24th international conference on Large installation system administration pp. 1-15.
38. Riak (2017) Riak.
39. Bappalige SP (2014) Hadoop
40. Sense Q (2017) Qlik Sense Desktop.
41. Sharma S, Mangat V (2015) Technology and Trends to Handle Big Data: Survey. In Fifth International Conference on Advanced Computing & Communication Technologies pp. 266-271.
42. Terrastore (2009) Terrastore.
43. Verborgh R, Wilde MDe (2013) Using OpenRefine. Packt Publishing.
44. Vijayaraj J, Saravanan R, Victer Paul P Raju R (2016) A comprehensive survey on big data analytics tools. Online International Conference on Green Engineering and Technologies (IC-GET) pp. 1-6.
45. Weave (2017) Weave.
46. Yu S, Liu M, Dou W, Liu X, Zhou S, et al. (2017) Networking for Big Data : A Survey, 19: 531-549.
47. Zoho (2017) Zoho Reports.
48. Cloudera (2017) Cloudera.
49. Sivarajah U, Kamal MM, Irani Z, Weerakkody V (2017) Critical analysis of Big Data challenges and analytical methods. Journal of Business Research. The Authors, 70: 263-286.
50. Loshin D (2014) Addressing Five Emerging Challenges of Big Data. p. 7
51. Norris D (2017) Hybrid Cloud Monitoring.
52. Wasson M (2017) Big data architecture style.
53. Hemsoth N. (2012) How 8 Small Companies are Retooling Big Data.
54. Prasad BR, Agarwal S (2016) Comparative Study of Big Data Computing and Storage Tools : A Review. International Journal of Database Theory and Application 9: 45-66.
55. Padgavankar MH Gupta SR (2014) Big Data Storage and Challenges. International Journal of Computer Science and Information Technologies 5: 2218-2223.
56. Bunnik A, Cawley A, Mulqueen M, Zwitter A (2016) Big data challenges: Society, security, innovation and ethics. Big Data Challenges: Society, Security, Innovation and Ethics.
57. Khan M, Wu X, Xu X, Dou W (2017) Big data challenges and opportunities in the hype of Industry 4.0. IEEE International Conference on Communications (ICC) 1-6.
58. ChartBlocks (2017) ChartBlocks.