Big Data Assignment: **Big Data Ecosystem**

Submitted by:

Name: Khom Raj Thapa Magar

Roll No.: 171347

Big Data Ecosystem

1. Introduction

Big Data is generally a combination of structured, semi-structured and unstructured data collected by organizations that can be mined for information and used in machine learning projects, predictive modeling and other advanced analytics applications.

According to Gartner,

“Big Data” is high-volume, velocity, and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

The volume dimension refers to the largeness of the data. The data size in a big data ecosystem can range from dozens of terabytes to a few zettabytes and is still growing.

The velocity dimension indicates the increasing speed at which big data is created and the increasing speed at which the data need to be stored and analyzed, while the variety dimension refers to increased diversity of data types.

Variety introduces additional complexity to data processing as more kinds of data need to be processed, combined and stored. While the 3 Vs have been continuously used to describe big data, further more Vs have been introduced.

Table 1. Big data characteristics

3 Vs	Volume	Vast amount of data that has to be captured, stored, processed and displayed
	Velocity	Rate at which the data is being generated, or analyzed
	Variety	Differences in data structure (format) or differences in data sources themselves (text, images, voice, geospatial data)
5 Vs	Veracity	Truthfulness (uncertainty) of data, authenticity, provenance, accountability
	Validity	Suitability of the selected dataset for a given application, accuracy and correctness of the data for its intended use
7 Vs	Volatility	Temporal validity and fluency of the data, data currency and availability, and ensures rapid retrieval of information as required
	Value	Usefulness and relevance of the extracted data in making decisions and capacity in turning information into action
10 Vs	Visualization	Data representation and understandability of methods (data clustering or using tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone trees)
	Vulnerability	Security and privacy concerns associated with data processing
	Variability	the changing meaning of data, inconsistencies in the data, biases, ambiguities, and noise in data

2. Big Data Ecosystem

The term Ecosystem is defined in scientific literature as a complex network or interconnected systems. With the advent of the web and cloud services, cloud computing is quickly displacing the traditional in-house system as a dependable, scalable, and affordable IT solution. Previously, organizations dealt with static, centrally stored data collected from numerous sources. Thus, large datasets – log files, social media sentiments, click-streams – are no longer expected to reside within a central server or within a fixed place in the cloud.

Big Data Ecosystem is the comprehension of massive functional components with different types of enabling tools. The advantages of its systematic platform and the potential of big data analytics are also part of the big data ecosystem's capabilities, which go beyond just computing and storing large amounts of data.

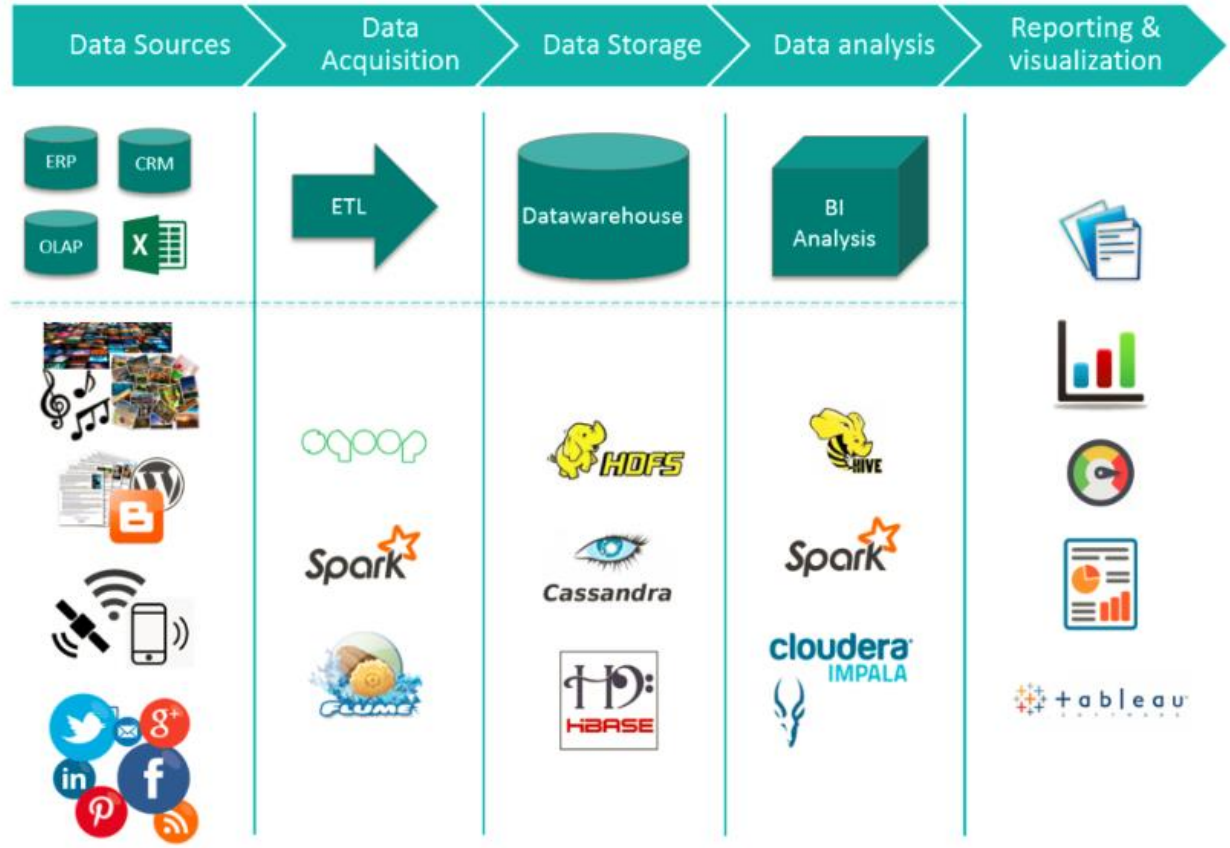

Figure 1. Big Data Ecosystem

The figure about depicts the diagram of a big data ecosystem. It comprises of components like Data Sources, Data Acquisition, Data Storage, Data Analysis and Reporting & visualization

Table 2. Examples of big data ecosystems

Facebook	Facebook (2018) has more than two billion users on millions of servers, running thousands of configurations changes every day involving trillions of configuration checks []
LinkedIn	It takes a lot of horsepower to support LinkedIn's 467 million members worldwide (in 2017), especially when you consider that each member is getting a personalized experience and a web page that includes only their contacts. Supporting the load are some 100,000 servers spread across multiple data centers []
Alibaba	The 402,000 web-facing computers that Alibaba hosts (2017) from China-allocated IP addresses would alone be sufficient to make Alibaba the second largest hosting company in the world today []
Google	There's no official data on how many servers there are in Google's data centers, but Gartner estimated in a July 2016 report that Google at the time had 2.5 million servers. Google data centers process an average of 40 million searches per second, resulting in 3.5 billion searches per day and 1.2 trillion searches per year, Internet Live Stats reports []
Amazon	... an estimate of 87 AWS datacenters in total and a range of somewhere between 2.8 and 5.6 million servers in Amazon's cloud (2014) []
Twitter	Twitter (2013) now has 150M worldwide active users, handles 300K queries per second (QPS) to generate timelines, and a firehose that churns out 22 MB/s. Some 400 million tweets a day flow through the system and it can take up to 5 min for a tweet to flow from Lady Gaga's fingers to her 31 million followers []

3. Components of Big Data Ecosystem

In order to depict the information flow in just a few phases, we can simply divide the processing workflow into three layers:

- Data sources
- Data management (integration, storage and processing)
- Data analytics, Business intelligence (BI) and knowledge discovery (KD)

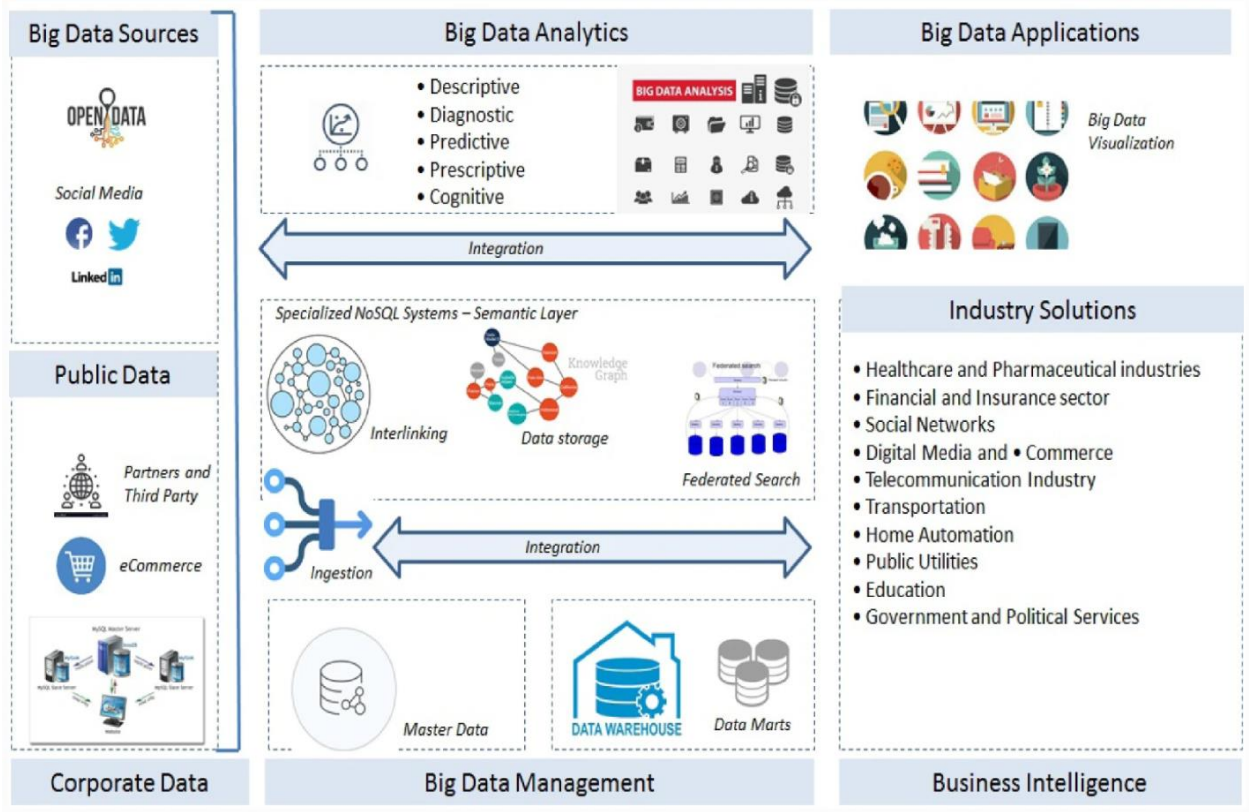

Figure 2. Components of big data ecosystem

3.1 Data Sources:

In a modern data ecosystem, the data source layer is composed of both private and public data sources. The corporate data originates from internal systems, cloud-based systems, as well as external data provided from partners and third parties.

Some of the examples of Open Data Sources from different domains are:

- **Facebook** Graph API, curated by Facebook, is the primary way for apps to read and write to the Facebook social graph. It is essentially a representation of all information on Facebook now and in the past.

- **Open Corporates** is one of the largest open datasets of companies in the world and holds hundreds of millions of datasets in essentially any country.
- **Global Financial Data**'s API is recommended for analysis who require large amounts of data for broad research needs. It enables researchers to study the interaction between different data series, sectors, and genres of data. The API supports R and Python so that the data can be directly uploaded to the target application.
- **Open Street Map** is a map of the world, created by people free to use under an open license. It powers map data on thousands of websites, mobile apps, and hardware devices.
- **The National Centers for Environmental Information** (NCEI) is responsible for hosting and providing access to one of the most significant archives on Earth, with comprehensive oceanic, atmospheric, and geophysical data.
- **DBPedia** is a semantic version of Wikipedia. It has helped companies like Apple, Google, and IBM to support artificial intelligence projects. DBPedia is in the center of the Linked Data cloud.

3.2 Data Management:

As data become increasingly available (from social media, web logs, IoT sensors etc.), the challenge of managing (selecting, combining, storing) and analyzing large and growing data sets is growing more urgent. From a data analytics point of view, that means that data processing has to be designed taking into consideration the diversity and scalability requirements of targeted data analytics applications.

Over the last two decades, the emerging challenges in the design of end-to-end data processing pipelines were addressed by computer scientists and software providers in the following ways:

- In addition to operational database management systems (present on the market since 1970s), different **NoSQL stores** appeared that lack adherence to the time-honored SQL principles of ACID (atomicity, consistency, isolation, durability).
- **Cloud Computing** emerged as a paradigm that focuses on sharing data and computations over a scalable network of nodes including end user computers, data centers, and web services.
- The **Data Lake** concept as a new storage architecture was promoted where raw data can be stored regardless of source, structure and (usually) size. The *data warehousing* approach (based on a repository of structured, filtered data that has already been processed for a specific purpose) is thus perceived as outdated as it creates certain issues with respect to data integration and the addition of new data sources.

3.3 Data Analytics:

Data analytics refers to technologies that are grounded mostly in data mining and statistical analysis. The selection of an appropriate processing model and analytical solution is a challenging problem and depends on the business issues of the targeted domain, for instance e-commerce

management, market intelligence, e-government, healthcare, energy efficiency, emergency management, production management, and/or security. Depending on the class of problem that is being solved (e.g. risk assessment in banks and the financial sector, predictive maintenance of wind farms, sensing and cognition in production plants, automatic response in control rooms, etc.), the *data analytics* solution also relies on text/web/network/mobile analytical services. Here various machine learning techniques such as association rule mining, decision trees, regression, support vector machines, and others are used.

While simple reporting and business intelligence applications that generate aggregated measurements across different predefined dimensions based on the data-warehousing concept were enough in 1990s, since 1995 the focus has been on introducing parallelism into machine learning.

4. Big Data Ecosystem Challenges

It would be a big mistake to believe that Big Data Ecosystem is not without its challenges. Big Data remains a new phenomenon. The Hadoop file system that underpins most current big data solutions has its roots in an antecedent dating from 2004. Hadoop permits the storage and access to large amounts of data, which in turn enables many analyses that were previously prohibitively time consuming. For example, extracting information from billions of records in order to identify clusters of related activity can be performed in minutes using the parallel computing capabilities of Hadoop and MapReduce rather than the days previously required.

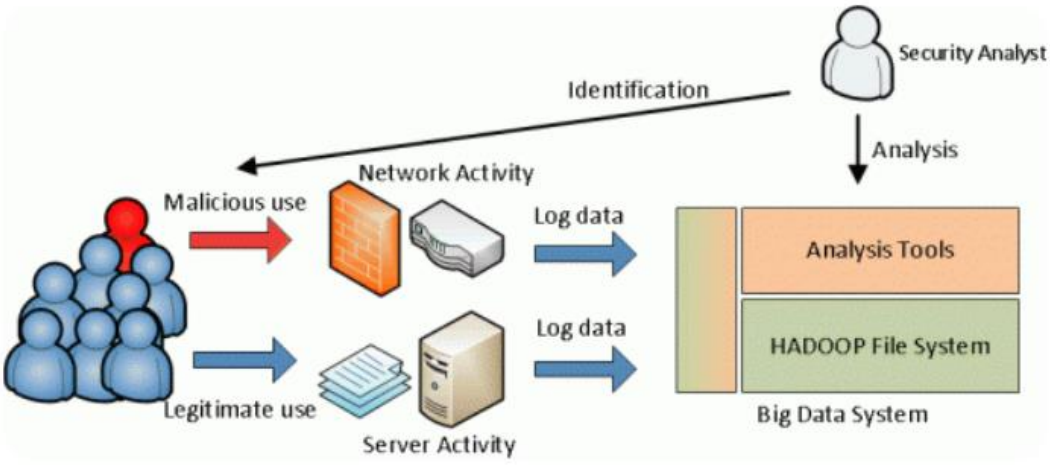

Figure 3. Big Data Ecosystem Security Challenges

The big data systems allow security analysts to collect log data to identify malicious activity.

Implementation of big data may be held back outdated attitudes and stale received wisdom as managers seek to apply old ways of doing things in a new world. Hype and false expectations may lead big data projects to be perceived as failing. Organizations must be realistic about what big data can deliver and what can be achieved. Working step by step toward a well-defined goal of addressing a small number of pressing business issues may be the best approach.

Big data is being used to transform many different disciplines, from improving health care, to reducing carbon emissions, to improving marketing efficiencies. However, each domain is different and techniques developed for one may not be easily portable to another. Skilled individuals will emerge over time, as will the development of custom data analysis tools. But for the moment we are often faced with having access to domain experts without data analysis skills.

Customizing tools and analytic algorithms is part and parcel of the current big data ecosystem. Intellectual freedom is part of the attraction for many researchers working in the field of big data. We can expect the ecosystem to mature over time, but for the moment it retains a certain "Wild-West" cachet for practitioners. The choice of approach to analyzing data is often dictated more by gut instinct and familiarity with certain algorithms.

Managing the resources of a big data cluster poses a large problem for system administrators. Enormous efficiencies can be leveraged by running big data queries in parallel on many machines. Administrative tools to manage this process are developing but may not yet be optimal in scheduling work within large clusters.

Sharing tasks and data across a large number of machines is one of the fundamental tenets of big data. This allows new machines to be added to a cluster to expand the disk storage and processing capacity as needed. System admins need to ensure data is stored in such a way that if a disk drive fails data is not lost. Big data systems may contain far too much data to be backed up on traditional media. Ensuring that there is no single point of failure within the cluster is vital.

Big data is fundamentally changing our approach to solving problems within information security. The success of the technique relies on a whole ecosystem of supporting tools, technologies and skilled individuals. As the technology develops, invariably the ecosystem will lag, leading to areas where supporting tools and relevant approaches are lacking.

5. Conclusion

Coming to the conclusion, Big Data Ecosystem is the comprehension of massive functional components with different types of enabling tools. Thus, this ecosystem can be grouped into technologies that have similar goals and functionalities.