UNIT-3

BIG DATA ANALYTICS

What is Big Data?

Data which are very large in size is called Big Data. Normally we work on data of size MB(WordDoc ,Excel) or maximum GB(Movies, Codes) but data in Peta bytes i.e. 10^15 byte size is called Big Data. It is stated that almost 90% of today's data has been generated in the past 3 years.

Sources of Big Data

These data come from many sources like,

Social networking sites: Facebook, Google, LinkedIn all these sites generates huge amount of data on a day to day basis as they have billions of users worldwide.

E-commerce site: Sites like Amazon, Flipkart, Alibaba generates huge amount of logs from which users buying trends can be traced.

Weather Station: All the weather station and satellite gives very huge data which are stored and manipulated to forecast weather.

Telecom company: Telecom giants like Airtel, Vodafone study the user trends and accordingly publish their plans and for this they store the data of its million users.

Share Market: Stock exchange across the world generates huge amount of data through its daily transaction.

3Vs of Big Data

Velocity: The data is increasing at a very fast rate. It is estimated that the volume of data will double in every 2 years.

Variety: Now a days data are not stored in rows and column. Data is structured as well as unstructured. Log file, CCTV footage is unstructured data. Data which can be saved in tables are structured data like the transaction data of the bank.

Volume: The amount of data which we deal with is of very large size of Peta bytes.

Big Data Analytics

The volume of data that one has to deal has exploded to unimaginable levels in the past decade, and at the same time, the price of data storage has systematically reduced.

Private companies and research institutions capture terabytes of data about their users' interactions, business, social media, and also sensors from devices such as mobile phones and automobiles. The challenge of this era is to make sense of this sea of data. This is where big data analytics comes into picture.

Big Data Analytics largely involves collecting data from different sources, munge it in a way that it becomes available to be consumed by analysts and finally deliver data products useful to the organization business.

The process of converting large amounts of unstructured raw data, retrieved from different sources to a data product useful for organizations forms the core of Big Data Analytics.

Industry Examples – Using Big Data

There are some major industries reshaping their functions using Big Data. They are,

1. Media and entertainment
2. Finance/Banking
3. Healthcare
4. Education
5. Retail
6. Manufacturing

Architectures, Frameworks, and Tools

The conceptual framework for a big data analytics project is similar to that for a traditional business intelligence or analytics project. The key difference lies in how the processing is executed. In a regular analytics project, the analysis can be performed with a business intelligence tool installed on a stand-alone system such as a desktop or laptop.

Since the big data is large by definition, the processing is broken down and executed across multiple nodes. While the concepts of distributed processing are not new and have existed for decades, their use in analyzing very large data sets is relatively new as companies start to tap into their data repositories to gain insight to make informed decisions.

Additionally, the availability of open-source platforms such as Hadoop/MapReduce on the cloud has further encouraged the application of big data analytics in various domains. Third, while the algorithms and models are similar, the user interfaces are entirely

different at this time. Classical business analytics tools have become very user-friendly and transparent.

On the other hand, big data analytics tools are extremely complex, programming intensive, and need the application of a variety of skills.

The following figure indicates, a primary component is the data itself. The data can be from internal and external sources, often in multiple formats, residing at multiple locations in numerous legacy and other applications. All this data has to be pooled together for analytics purposes.

The data is still in a raw state and needs to be transformed. Here, several options are available. A service-oriented architectural approach combined with web services (middleware) is one possibility. The data continues to be in the same state, and services are used to call, retrieve, and process the data.

On the other hand, data warehousing is another approach wherein all the data from the different sources are aggregated and made ready for processing. However, the data is unavailable in real time. Via the steps of extract, transform, and load (ETL), the data from diverse sources is cleansed and made ready. Depending on whether the data is structured or unstructured, several data formats can be input to the Hadoop/MapReduce platform.

In this next stage in the conceptual framework, several decisions are made regarding the data input approach, distributed design, tool selection, and analytics models. Finally, to the far right the four typical applications of big data analytics are queries, reports, online analytic processing (OLAP), and data mining.

Visualization is an overarching theme across the four applications. A wide variety of techniques and technologies have been developed and adapted to aggregate, manipulate, analyze, and visualize big data. These techniques and technologies draw from several fields, including statistics, computer science, applied mathematics, and economics.

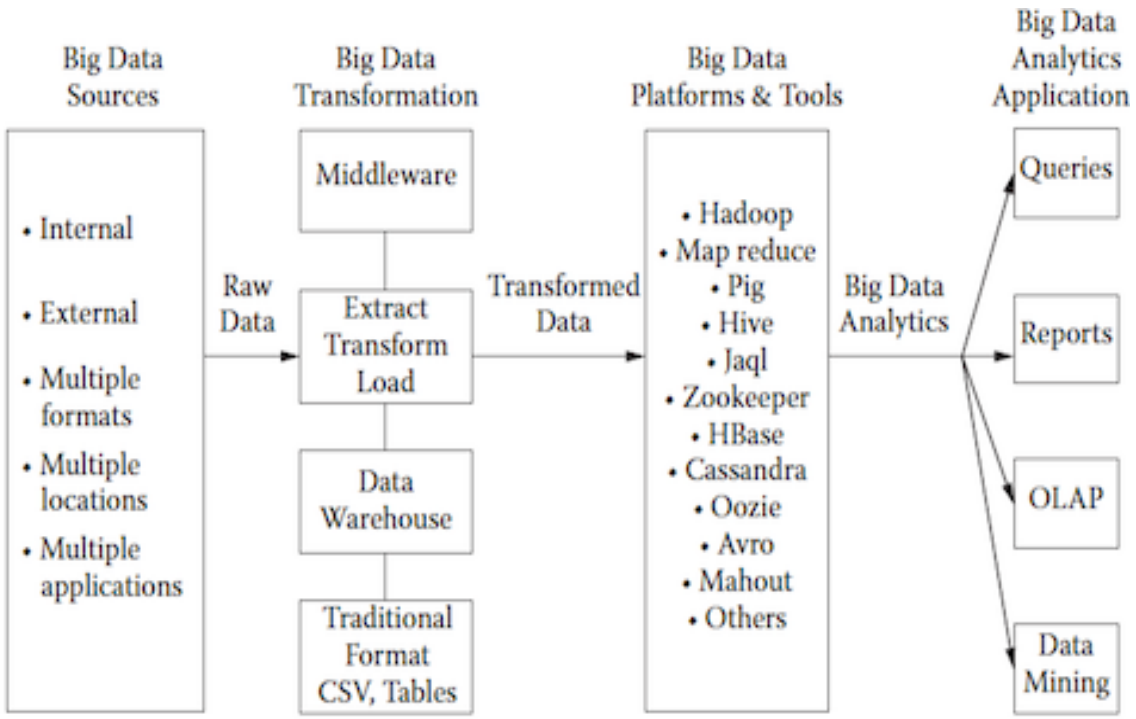

Fig.(1): An applied conceptual architecture of big data analytics.

Big Data Tools

Today's market is flooded with an array of Big Data tools and technologies. They bring cost efficiency, better time management into the data analytical tasks. Some of the best Big data tools are,

1. Hadoop
2. HPCC
3. Storm
4. Qubole
5. Cassandra
6. Statwing
7. CouchDB
8. Pentaho
9. Flink
10. Cloudera
11. OpenRefine
12. RapidMiner
13. DataCleaner
14. Kaggle
15. Hive

Challenges of Big Data Analytics

There are several challenges that can impede risk managers' ability to collect and use analytics. Some of the major challenges that big data analytics program are facing today include the following:

Uncertainty of Data Management Landscape: Because big data is continuously expanding, there are new companies and technologies that are being developed every day. A big challenge for companies is to find out which technology works bests for them without the introduction of new risks and problems.

The Big Data Talent Gap: While Big Data is a growing field, there are very few experts available in this field. This is because Big data is a complex field and people who understand the complexity and intricate nature of this field are far few and between. Another major challenge in the field is the talent gap that exists in the industry

Getting data into the big data platform: Data is increasing every single day. This means that companies have to tackle a limitless amount of data on a regular basis. The scale and variety of data that is available today can overwhelm any data practitioner and that is why it is important to make data accessibility simple and convenient for brand managers and owners.

Need for synchronization across data sources: As data sets become more diverse, there is a need to incorporate them into an analytical platform. If this is ignored, it can create gaps and lead to wrong insights and messages.

Getting important insights through the use of Big data analytics: It is important that companies gain proper insights from big data analytics and it is important that the correct department has access to this information. A major challenge in big data analytics is bridging this gap in an effective fashion.

Big Data Analytics in Healthcare

Big data has changed the way we manage, analyze, and leverage data across industries. One of the most notable areas where data analytics is making big changes is healthcare.

In fact, healthcare analytics has the potential to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases, and improve the quality of life in general. The average human lifespan is increasing across the world population, which poses new challenges to today's treatment delivery methods.

Big data in healthcare is a term used to describe massive volumes of information created by the adoption of digital technologies that collect patients' records and help in

managing hospital performance, otherwise too large and complex for traditional technologies.

The application of big data analytics in healthcare has a lot of positive and also life-saving outcomes. In essence, big-style data refers to the vast quantities of information created by the digitization of everything that gets consolidated and analyzed by specific technologies. Applied to healthcare, it will use specific health data of a population (or of a particular individual) and potentially help to prevent epidemics, cure disease, cut down costs, etc.

Now that we live longer, treatment models have changed and many of these changes are namely driven by data. Doctors want to understand as much as they can about a patient and as early in their life as possible, to pick up warning signs of serious illness as they arise – treating any disease at an early stage is far simpler and less expensive. By utilizing key performance indicators in healthcare and healthcare data analytics, prevention is better than cure, and managing to draw a comprehensive picture of a patient will let insurance provide a tailored package.

This is the industry's attempt to tackle the siloes problems a patient's data has: everywhere are collected bits and bytes of it and archived in hospitals, clinics, surgeries, etc., with the impossibility to communicate properly.

Indeed, for years gathering huge amounts of data for medical use has been costly and time-consuming. With today's always-improving technologies, it becomes easier not only to collect such data but also to create comprehensive healthcare reports and convert them into relevant critical insights that can then be used to provide better care.

This is the purpose of healthcare data analytics: using data-driven findings to predict and solve a problem before it is too late, but also assess methods and treatments faster, keep better track of inventory, involve patients more in their own health, and empower them with the tools to do so

Big Data Applications In Healthcare

1. Electronic Health Records(EHRs)
2. Telemedicine
3. Improved Strategic Planning
4. Prevent opioid abuse
5. Predictive Analytics
6. Reducing Fraud
7. Enhancing Patient Engagement
8. Developing new therapies
9. Optimizing ER admissions
10. Smart Staffing
11. Learning and Development
12. Risk & Diseases Management
13. Self-harm Prevention
14. Improved patients' Predictions
15. Improved supply-chain
16. Medical Imaging
17. Real-time alerting
18. Augment cancer treatment