SNS COLLEGE OF TECHNOLOGY
COIMBATORE-35

DEPARTMENT OF INFORMATION TECHNOLOGY

19ITE305 – BIG DATA ANALYTICS

UNIT I: INTRODUCTION TO BIG DATA AND ANALYTICS
Topic 6: Challenges - Big Data Analytics important - Data Science - Data Scientist

CHALLENGES OF BIG DATA

There are mainly seven challenges of big data: scale, security, schema, Continuous availability, Consistency, Partition tolerant and data quality.

- Scale: Storage (RDBMS (Relational Database Management System) or NoSQL (Not only SQL)) is one major concern that needs to be addressed to handle the need for scaling rapidly and elastically. The need of the hour is a storage that can best withstand the attack of large volume, velocity and variety of big data. Should you scale vertically or should you scale horizontally?
- Security: Most of the NoSQL big data platforms have poor security mechanisms (lack of proper authentication and authorization mechanisms) when it comes to safeguarding big data. A spot that cannot be ignored given that big data carries credit card information, personal information and other sensitive data.
- schema: Rigid schemas have no place. We want the technology to be able to fit our big data and not the other way around. The need of the hour is dynamic schema. Static (pre-defined schemas) are obsolete.
- Continuous availability: The big question here is how to provide 24/7 support because almost all RDBMS and NoSQL big data platforms have a certain amount of downtime built in.
- Consistency: Should one opt for consistency or eventual consistency? Partition tolerant: How to build partition tolerant systems that can take care of both hardware and software failures?
- Data quality: How to maintain data quality- data accuracy, completeness, timeliness, etc.? Do we have appropriate metadata in place?

Importance of Big Data Analytics

help us with? It allows the businesses to make faster and better decisions by providing the right information to the right person at the right time in 23 the right format.

It is about analysis of the past or historical data and then displaying the findings of the analysis or reports in the form of enterprise dashboards, alerts, notifications, etc. It has support for both pre-specified reports as well as ad hoc querying.

Reactive - Big Data Analytics: Here the analysis is done on huge datasets but the approach is still reactive as it is still based on static data.

Proactive - Analytics: This is to support futuristic decision making by use of data mining predictive modelling, text mining, and statistical analysis on. This analysis is not on big data as it still the traditional database management practices on big data and therefore has severe limitations on the storage capacity and the processing capability.

Proactive - Big Data Analytics: This is filtering through terabytes, petabytes, exabytes of information to filter out the relevant data to analyze. This also includes high performance analytics to gain rapid insights from big data and the ability to solve complex problems using more data.

Big Data Technologies

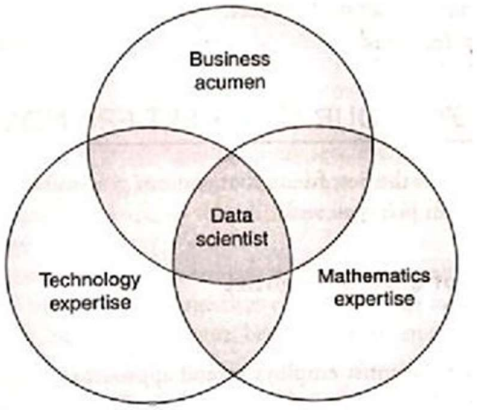
Following are the requirements of technologies to meet challenges of big data:

- The first requirement is of cheap and ample storage.
- We need faster processors to help with quicker processing of big data. Affordable open source distributed big data platforms, such as Hadoop.
- Parallel processing, clustering, virtualization, large grid environments (to distribute processing to a number of machines), high connectivity, and high throughputs(rate at which something is processed). Cloud computing and other flexible resource allocation arrangements.

Data Science

Data science is the science of extracting knowledge from data. In other words, it is a science of drawing out hidden patterns amongst data using statistical and mathematical techniques. It employs techniques and theories drawn from many fields from the broad areas of mathematics, statistics, information technology including machine learning, data engineering, probability

models, statistical learning, pattern recognition and learning, etc. Data Scientist works on massive datasets for weather predictions, oil drillings, earthquake prediction, financial frauds, terrorist network and activities, global economic impacts, sensor logs, social media analytics, customer churn, collaborative filtering (prediction about interest on users), regression analysis, etc. Data science is multi-disciplinary



Business Acumen(expertise) Skills

A data scientist should have following ability to play the role of data scientist.

- Understanding of domain
- Business strategy
- Problem solving
- Communication
- Presentation
- Keenness

Technology Expertise

Following skills required as far as technical expertise is concerned.

- Good database knowledge such as RDBMS.
- Good NoSQL database knowledge such as MongoDB, Cassandra, HBase, etc.
- Programming languages such as Java. Python, C++, etc.
- Open-source tools such as Hadoop.
- Data warehousing.
- Data mining
- Visualization such as Tableau, Flare, Google visualization APIs, etc.

Mathematics Expertise

The following are the key skills that a data scientist will have to have to comprehend data, interpret it and analyze.

- Mathematics.
- Statistics.
- Artificial Intelligence (AI).
- Algorithms.
- Machine learning.
- Pattern recognition.
- Natural Language Processing.
- To sum it up, the data science process is
- Collecting raw data from multiple different data sources.
- Processing the data.
- Integrating the data and preparing clean datasets.
- Engaging in explorative data analysis using model and algorithms.
- Preparing presentations using data visualizations.
- Communicating the findings to all stakeholders.
- Making faster and better decisions.