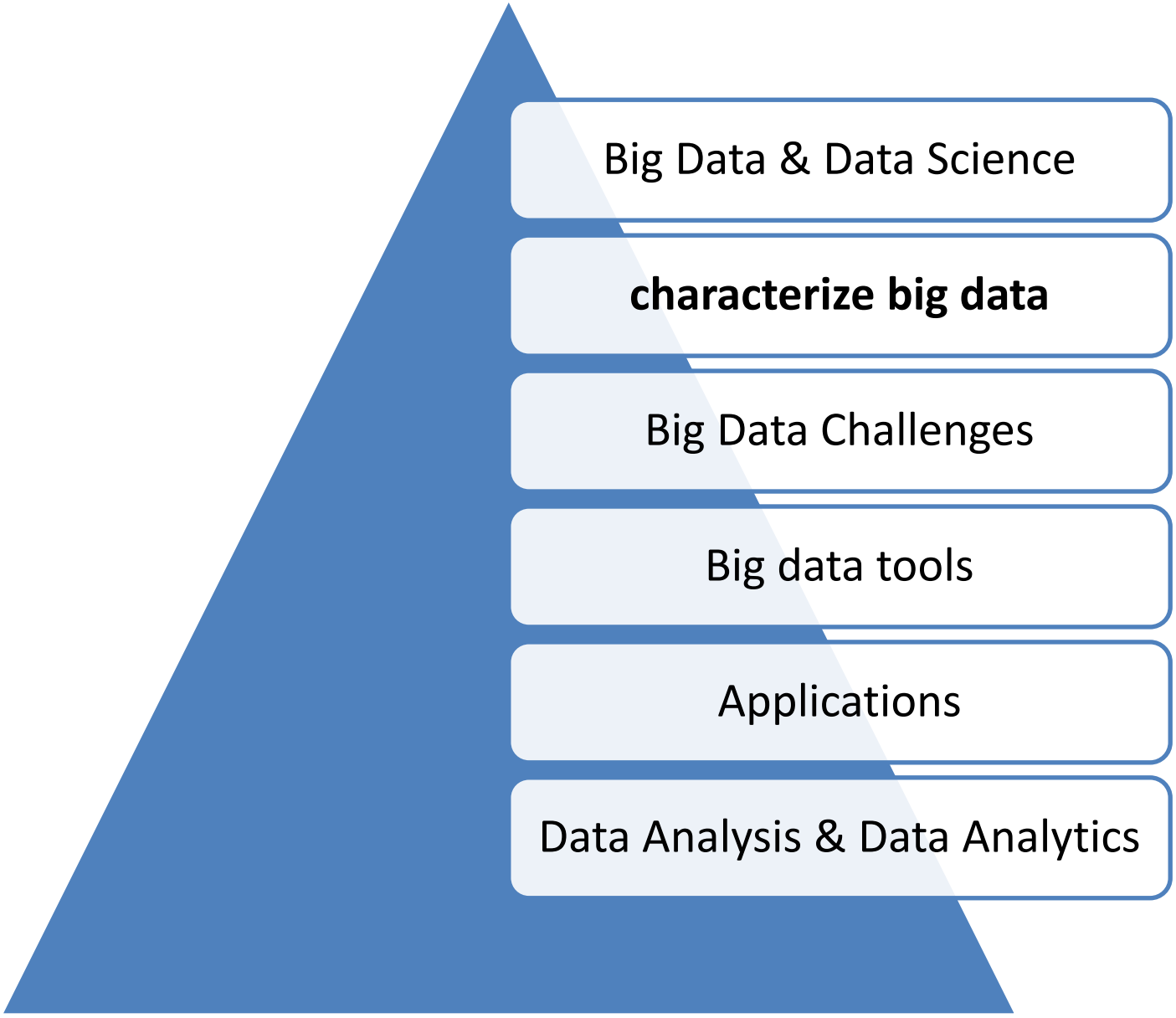
Introduction to Bigdata and Hadoop

Lecture 1

Dr. S. Srivastava


Big Data & Data Science
characterize big data
Big Data Challenges
Big data tools
Applications
Data Analysis & Data Analytics

Big Data & Data Science

- lots and lots of web pages …
- a billion Facebook users
- billion+ Facebook pages
- hundreds of million Twitter accounts
- hundreds of million tweets per day
- Billions of Google queries per day
- Millions of servers, petabytes of data

In contrast, typical large enterprise:

- 5000-50,000 servers,
- Terabytes of data, millions of Txn/day

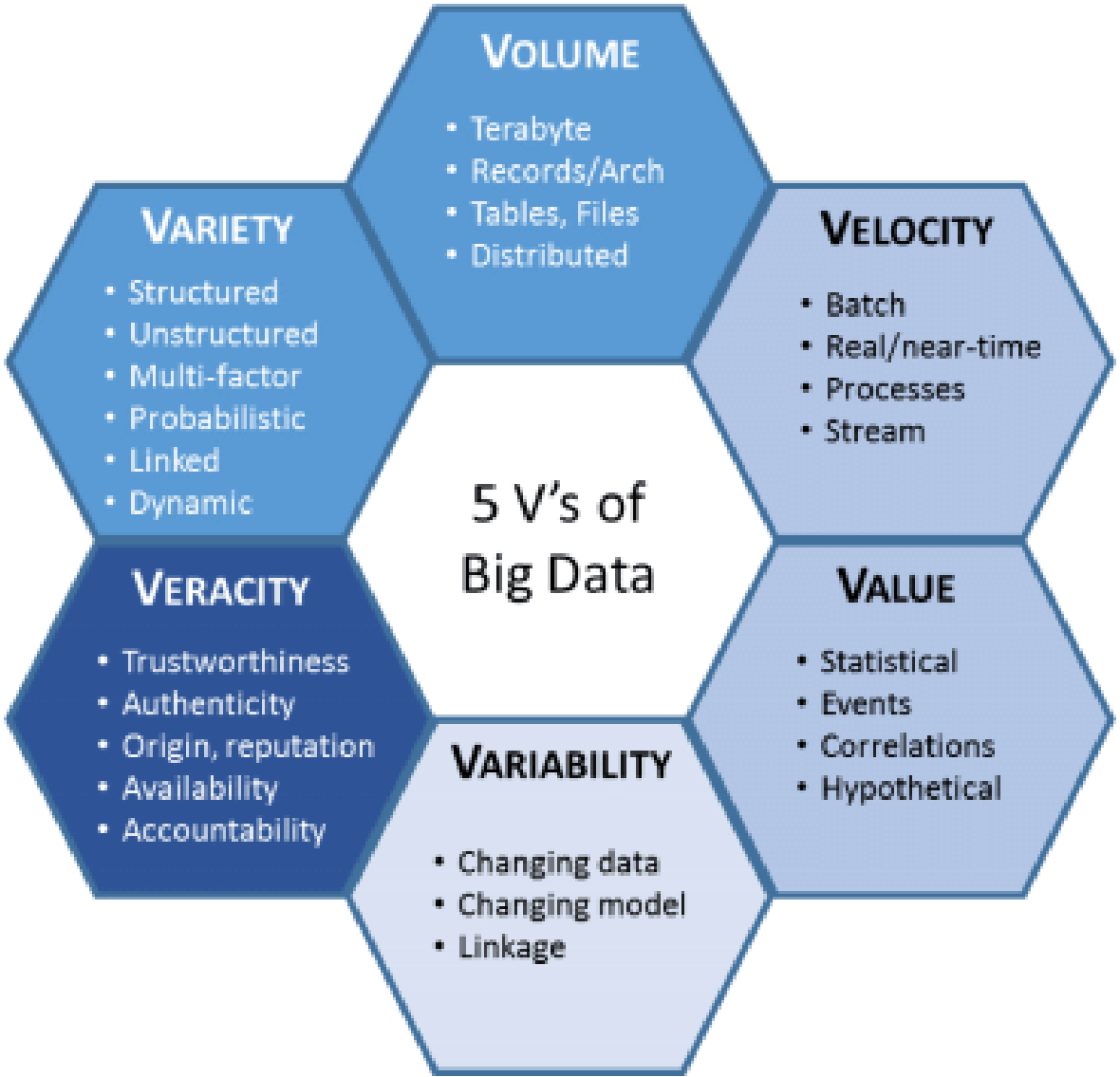
Big data is a blanket term for any collection of data sets so large or complex that it becomes difficult to process them using traditional data management techniques such as, for example, the RDBMS (relational database management systems).

The widely adopted RDBMS has long been regarded as a one-size-fits-all solution, but the demands of handling big data have shown otherwise.

Data science involves using methods to analyze massive amounts of data and extract the knowledge it contains.


Data Analysis from large data set
Knowledge extraction
BIG DATA

How can we characterize big data?



How can we characterize big data?

Big data is commonly characterized using a number of V's.

- The first three are **volume, velocity**, and **variety**.

(a) **Volume** refers to the vast amounts of data that is generated every second, minutes, hour, and day in our digitized world.

(b) **Variety** refers to the ever increasing different forms that data can come in such as text, images, voice, and geospatial data.

(c) **Velocity** refers to the speed at which data is being generated and the pace at which data moves from one point to the next.

Characteristics of Big Data

- **Veracity** refers to the biases, noise, and abnormality in data. Or, better yet, It refers to the often immeasurable uncertainties and truthfulness and trustworthiness of data.

- **Value** understand the costs and benefits of collecting and analyzing the data to ensure that ultimately the data that is gained can be monetized.

Challenges

1. Dealing with data growth

- storing and analyzing all that information.
- Much of that data is unstructured, meaning that it doesn't reside in a database.
- Documents, photos, audio, videos and other unstructured data can be difficult to search and analyze.

2. Generating insights in a timely manner

- Decreasing expenses through operational cost efficiencies
- Establishing a data-driven culture
- Creating new avenues for innovation and disruption
- Accelerating the speed with which new capabilities and services are deployed
- Launching new product and service offerings

3.Recruiting and retaining big data talent

- Job

4. Integrating disparate data sources

- Big data comes from a lot of different places — enterprise applications, social media streams, email systems, employee-created documents, etc.

5. Validating data

- Often organizations are getting similar pieces of data from different systems, and the data in those different systems doesn't always agree.
- The process of getting those records to agree, as well as making sure the records are accurate, usable and secure, is called data governance

6. Securing big data

- attractive targets for hackers or advanced persistent threats

7. Organizational resistance

- It is not only the technological aspects of big data that can be challenging — people can be an issue too
- Insufficient organizational alignment
- Lack of middle management adoption and understanding
- Business resistance or lack of understanding (41.0 percent)

BIG DATA TOOLS

1. Hadoop

- Apache Hadoop is the most prominent and used tool in big data industry with its enormous capability of large-scale processing data.
- This is 100% open source framework and runs on commodity hardware in an existing data center. Furthermore, it can run on a cloud infrastructure.
- Hadoop consists of four parts:
  - **Hadoop Distributed File System:** Commonly known as HDFS, it is a distributed file system compatible with very high scale bandwidth.
  - **MapReduce:** A programming model for processing big data.
  - **YARN:** It is a platform used for managing and scheduling Hadoop's resources in Hadoop infrastructure.
  - **Libraries:** To help other modules to work with Hadoop.

2. Apache Spark

- this open source big data tool is it fills the gaps of Apache Hadoop concerning data processing.
- Spark can handle both batch data and real-time data.
- It processes data much faster than traditional disk processing.
- Apache Spark is flexible to work with HDFS (Hadoop distributed file system) as well as with other data stores
- Spark Core is the heart of the project, and it facilitates many things like distributed task transmission, scheduling, I/O functionality

3. Apache Storm

- Apache Storm is a distributed real-time framework for reliably processing the unbounded data stream. The framework supports any programming language. The features of Apache Strom is as follows
- Massive scalability
- Fault-tolerance
- “fail fast, auto restart” approach
- The guaranteed process of every tuple
- Written in Clojure
- Runs on the JVM
- Supports direct acrylic graph(DAG) topology
- Supports multiple languages
- Supports protocols like (JavaScript Object Notation)

4. Cassandra

- distributed type database to manage a large set of data across the servers
- processes structured data sets
- Continuous availability as a data source
- Linear scalable performance
- Simple operations
- Across the data centers easy distribution of data
- Cloud availability points
- can handle numerous concurrent users across data centers

5. RapidMiner

- a software platform for data science activities and provides an integrated environment for:
- Preparing data
- Machine learning
- Text mining
- Predictive analytics
- Deep learning
- Application development
- Prototyping
- RapidMiner follows a client/server model where the server could be located on-premise, or in a cloud infrastructure.

6. MongoDB

- MongoDB is an open source NoSQL database which is cross-platform compatible with many built-in features.
- It runs on MEAN software stack, NET applications and, Java platform.
- It can store any type of data like integer, string, array, object, boolean, date etc.
- It provides flexibility in cloud-based infrastructure.
- It is flexible and easily partitions data across the servers in a cloud structure.
- MongoDB uses dynamic schemas. Hence, you can prepare data on the fly and quickly. This is another way of cost saving.

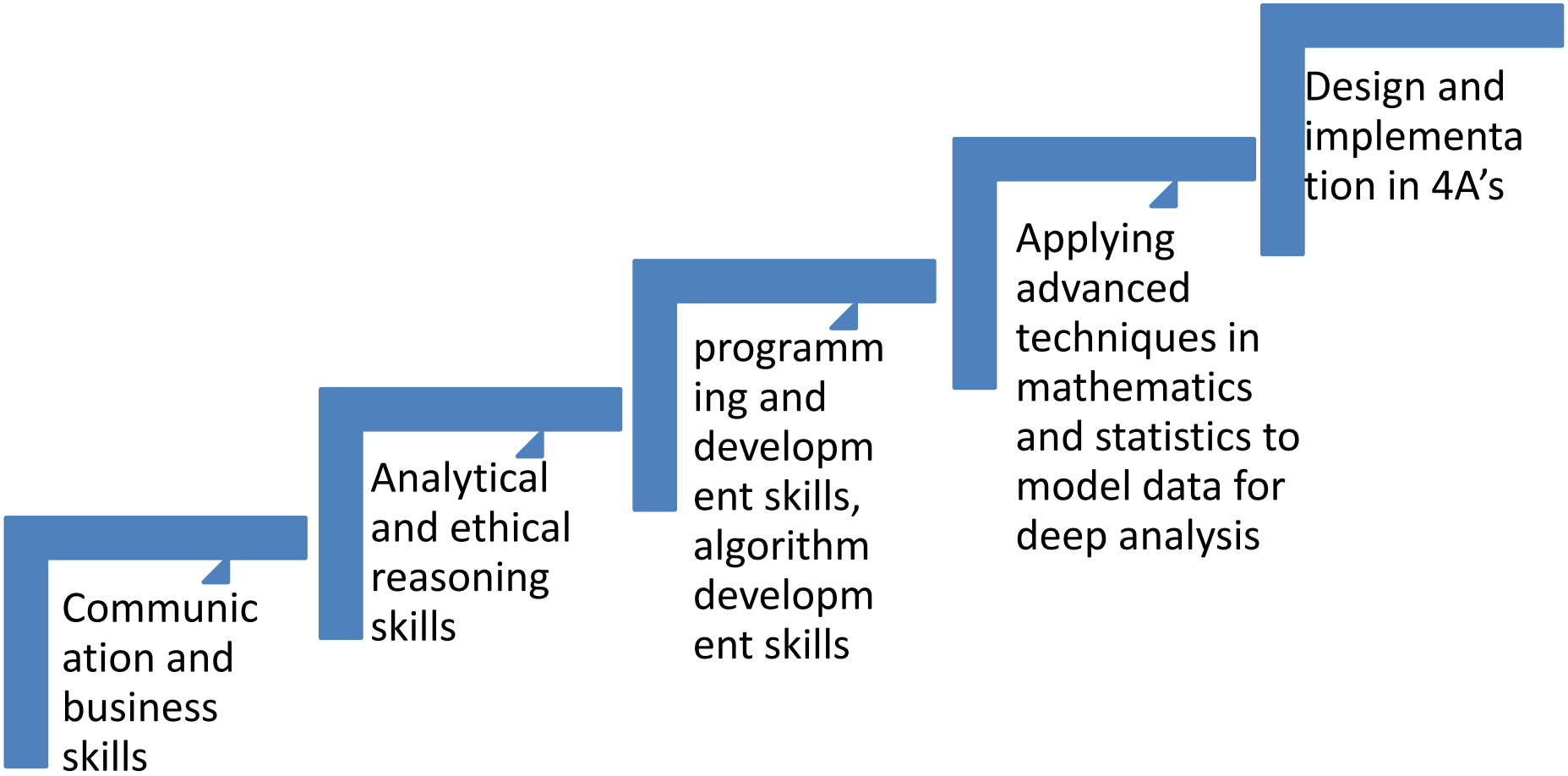
R Programming Tool

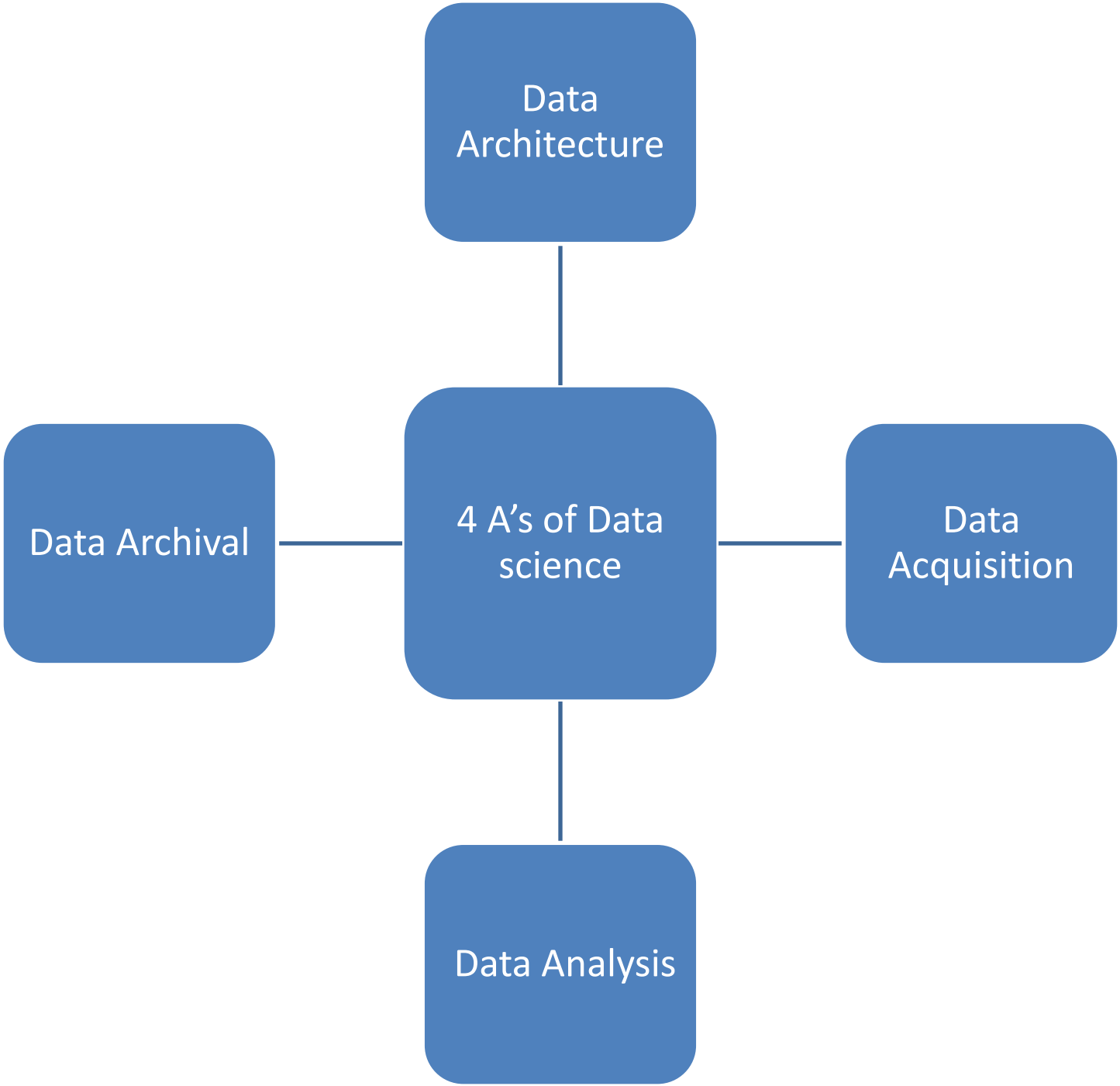
- although used for statistical analysis, as a user you don’t have to be a statistical expert.
- R has its own public library CRAN (Comprehensive R Archive Network) which consists of more than 9000 modules and algorithms for statistical analysis of data.
- R can run on Windows and Linux server as well inside SQL server. It also supports Hadoop and Spark.
- Using R tool one can work on discrete data and try out a new analytical algorithm for analysis.
- R model built and tested on a local data source can be easily implemented in other servers or even against a Hadoop data lake.

8. Neo4j

- Neo4j is one of the big data tools that is widely used graph database in big data industry. It follows the fundamental structure of graph database which is interconnected node-relationship of data.
- It supports ACID (Atomicity, Consistency, Isolation, Durability)transaction
- High availability
- Scalable and reliable
- Flexible as it does not need a schema or data type to store data
- It can integrate with other databases
- Supports query language for graphs which is commonly known as Cypher

Data Scientists


Communication and business skills
Analytical and ethical reasoning skills
programming and development skills, algorithm development skills
Applying advanced techniques in mathematics and statistics to model data for deep analysis
Design and implementation in 4A's


Data Architecture
Data Archival
4 A's of Data science
Data Acquisition
Data Analysis

#1. Main Task

Data Scientist

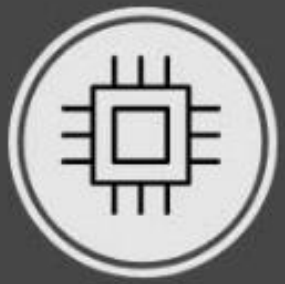
Ensure end to end flow of data lake architecture, starting from data loading till presentation to end user.

Big Data

Ensure huge data loading smoothly and fetching
those data for preparing big data dictionary which can be easily used for presenting end use by applying business rules.

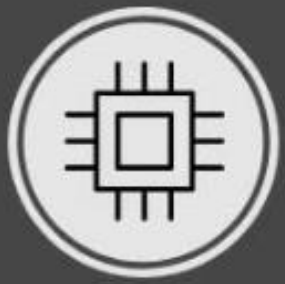
#2. Knowledge

Data Scientist



Should have knowledge of entire flow, including
business rules, current organization business track and user friendly presentation for end user.

Big Data



Should have knowledge on huge data loading
smoothly from various sources, and fetching data as quick as possible without any mistake.

#3. Technology

Data Scientist

Data Scientist normally have idea of all the technologies or processing tool like Hive, Map Reduce, R, Spark or the related technologies or tools.

Big Data

Those guys have clear ideas on data loading and data fetching related technologies or tools. There normally experts on Hive, Spark, MapReduce, Pig, Cassandra etc.

For further reading :

https://www.educba.com/data-scientist-vs-big-data/

Big Data Applications


Big Data
Crime Prediction and Prevention
Manufacturing
Media & Entertainment
Internet of Things (IoT)
Government
Pharmaceutical Drug Evaluation
Traffic Optimization
Weather Forecasting
Cyber security & Intelligence
Healthcare

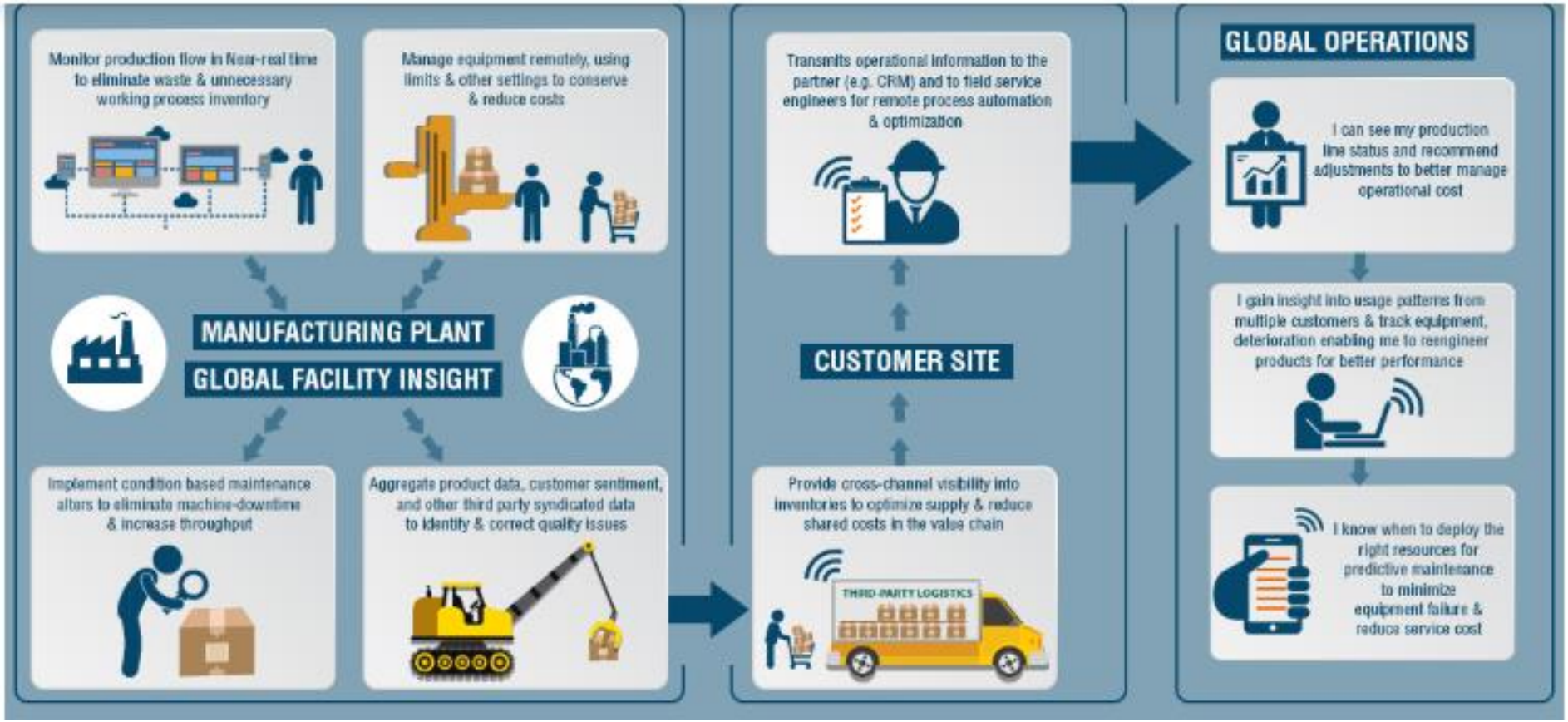
Big Data Applications: Healthcare

personalized medicine and prescriptive analytics.

identify patterns related to drug side effects, and gains other important information that can help patients and reduce costs.

wearable technologies that includes electronic health record data, imaging data, patient generated data, sensor data, and other forms of data.

Big Data Applications: Manufacturing



Product quality and defects tracking

Supply planning

Manufacturing process defect tracking

Output forecasting and increasing energy efficiency

Testing and simulation of new manufacturing processes

Support for mass-customization of manufacturing

Big Data Applications: Media & Entertainment

Predicting what the audience wants and scheduling optimization

Increasing acquisition and retention

Content monetization and new product development

Advertisement targeting

Big Data Applications: Internet of Things (IoT)

Data extracted from ***IoT*** devices provides a mapping of device inter-connectivity.

increasingly adopted as a means of gathering sensory data, and this sensory data is used in medical and manufacturing contexts.

Big Data Applications: Government

efficiencies in terms of cost, productivity, and innovation

the same data sets are often applied across multiple applications & it requires multiple departments to work in collaboration.

Government majorly acts in all the domains, thus it plays an important role in innovating Big Data applications in each and every domain.

Cyber security & Intelligence

The federal government launched a cyber security research and development plan that relies on the ability to analyze large data sets in order to improve the security of U.S. computer networks.

The National Geospatial-Intelligence Agency is creating a “Map of the World” that can gather and analyze data from a wide variety of sources such as satellite and social media data.

Crime Prediction and Prevention

real-time analytics to provide actionable intelligence that can be used to understand criminal behaviour, identify crime/incident patterns, and uncover location-based threats.

Weather Forecasting

The NOAA (National Oceanic and Atmospheric Administration) gathers data every minute of every day from land, sea, and space-based sensors. Daily NOAA uses Big Data to analyze and extract value from over 20 terabytes of data.

Tax Compliance

analyze both unstructured and structured data from a variety of sources in order to identify suspicious behavior and multiple identities. This would help in tax fraud identification.

Traffic Optimization

real-time traffic data gathered from road sensors, GPS devices and video cameras.

The potential traffic problems in dense areas can be prevented by adjusting public transportation routes in real time.

The **difference between data analysis and data analytics** is that **data analytics** is a broader term of which **data analysis** forms a subcomponent.

Data analysis refers to the process of compiling and analysing **data** to support decision making, whereas**data analytics** also includes the tools and techniques use to do so.