INTRODUCTION TO BIG DATA
STRUCTURED AND UNSTRUCTURED DATA
RELEVANCE OF BIG DATA IN AI

Fundamentals of Artificial Intelligence

Session 15

Pramod Sharma
pramod.sharma@prasami.com

2

Agenda

- Introduction to Data Analytics
- Type of Data
- Introduction to Big Data
- Hadoop Eco System
- Tasks of AI Using Big Data

pra-sâmi

3

"Social Media Analytics is all about mining retail-related insights from social channels, a perilous and personally exciting task to us. When our team spent the 22nd November feverishly following the social retail pulse on Black Friday, we knew the world wasn't preparing for an apocalypse." - Arun Prasath, Principal Engineer, WalmartLabs

pra-sâmi

4

Data Analytics

What?

- It is a process of transforming processed Information into Knowledge

Why do we need it?

- To predict progress, increase profit and for efficient utilization of our resources

Where do we need it?

- Everywhere:
  - Daily Life: buying the grocery, filling the fuel, online shopping, manage AC in home, tracking health records
  - Business : To increase Sales
  - Computer System : Various algorithms (LRU, MRU , Command Queuing Algorithms)

pra-sâmi

5

Interconnected world

- Once you search a product, on a website or in an app, you are shown advertisements of similar products in:
  - Apps,
  - Promotional mails on Yahoo and Gmail
  - Advertisements in Facebook, Instagram, YouTube, Pintrest
  - Yahoo mail side bar, web pages
  - Advertisements on Television - Binge, Netflix, Amazon Prime, etc.
- In some countries, even if you mention a product over phone to a friend
- Ever wondered?

pra-sâmi

7

DATA ANALYTICS

Month	Jul	Aug	Sep	Oct	Nov	Dec
Profit (millions)	2.0	2.1	2.2	2.1	2.3	2.4

pra-sâmi

8

DATA ANALYTICS

Month	Jul	Aug	Sep	Oct	Nov	Dec
Profit (millions)	2.0	2.1	2.2	2.1	2.3	2.4

pra-sâmi

9

Businesses Need Support for Decision Making

- Uncertain economics
- Rapidly changing environments
- Global competition
- Demanding customers
- Taking advantage of information acquired by companies is a Critical Success Factor.

pra-sâmi

10

The Information Gap

- The shortfall between gathering information and using it for decision making
  - Organizations have inadequate data warehouses

Data Warehousing Institute

Business Analysts spend 2 days a week gathering and formatting data, instead of performing analysis

- Business Intelligence (BI) seeks to bridge the information gap

pra-sâmi

11

Analysis Needs Data

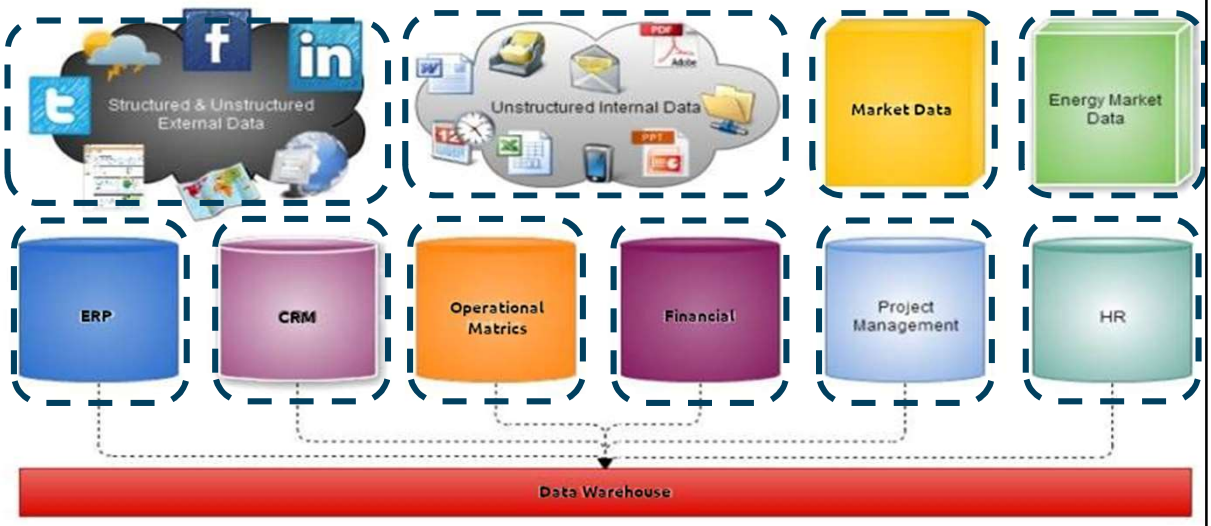
So Needs to Understand Various Types of DATA Over Which Companies Work

- Structured Data
- Semi Structured Data
- Unstructured
- BIG DATA

pra-sâmi

12

Enterprise Data Warehouse or Data Lake



13

Structured Data

- All data which can be stored in database in table with rows and columns
- They have relational key and can be easily mapped into pre-designed fields
- Today, this data are the most processed in development and the simplest way to manage information
- It has Data Dictionary (Glossary Terms)
- It is under strict Data Governance

14

Semi Structured Data

- Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it somewhat easy to analyze
- With some processing, can be stored in relation database eventually
  - Old school of thoughts
- We not need to store cleaned up version, we only need the information it may contain
  - Storing could be very hard for some of the semi structured data
  - But the semi structure exist to ease space, clarity or compute…
- Examples :
  - CSV , XML, JSON
- Structured data, semi structured data represents a small parts of data (5 to 10%).

15

Unstructured data

- Unstructured data represent around 80% of data
- It often include text and multimedia content
- Examples include
  - e-mail messages, Social Media messages, Twitter feeds, Comments on YouTube videos, Videos, photos, audio files, Presentations, Webpages and many other kinds of business documents
- Unstructured data is everywhere
- In fact, most individuals and organizations conduct their lives around unstructured data
- Just as with structured data, unstructured data is either machine generated or human generated

16

Unstructured data

- Here are some examples of machine-generated unstructured data:
  - IOT data: Every time you apply breaks in your car, it generates more than 10,000 data points
  - Satellite images: Weather data, Satellite surveillance imagery
    - Think about Google Earth
  - Scientific data: Seismic imagery, atmospheric data, and high energy physics
  - Photographs and video: Security, surveillance, traffic surveillance etc.
  - Radar or sonar data: Vehicular, meteorological, oceanographic & seismic profiles, etc.
- The following list shows a few examples of human-generated unstructured data:
  - Social media data: YouTube, Facebook, Twitter, LinkedIn, Instagram, etc.
  - Mobile data: Text messages, location information, physical movement, etc.
  - Website: Web logs, click streams, etc.

17

Business Analytics, BI, Big Data, Data Mining - What's the difference?

- Business Analytics – Tools to explore past data to gain insight into future business decisions
- BI – Tools and techniques to turn data into meaningful information
- Big Data –data sets that are so large or complex that traditional data processing applications are inadequate
- Data Mining - Tools for discovering patterns in large data sets

18

Hardware

Traditions

- Exotic HW
  - Big central servers
  - SAN – Storage Area Network
  - RAID – Redundant Arrays of Independent Disks
- Reliable Hardware
- Expensive scalability
- Expensive maintenance

Big Data

- Commodity HW
  - Rack of “pizza” boxes
  - Connected with Ethernet
  - JBOD – Just a Bunch of Disks
- Unreliable HW
  - built-in redundancy
- Cheaper to scale
- Competitive maintenance
- Overall cost effective

pra-sâmi

19

Big Data and AI

pra-sâmi

20

Big Data and Artificial Intelligence

- The world was entrenched in big data before it even realized that big data existed
- By the time the term was coined in 1998, big data had accumulated a massive amount of stored information
- Proper analysis could reveal valuable insights into the industry to which that particular data belonged
- To use the data. we need to
  - Sifting or sieve it
  - Parse it : converting it into a format more easily understood
  - Analyze it: to improve business decision-making processes
- All that is too much for human minds to tackle
- AI algorithms are needed to derive insight out of complex data

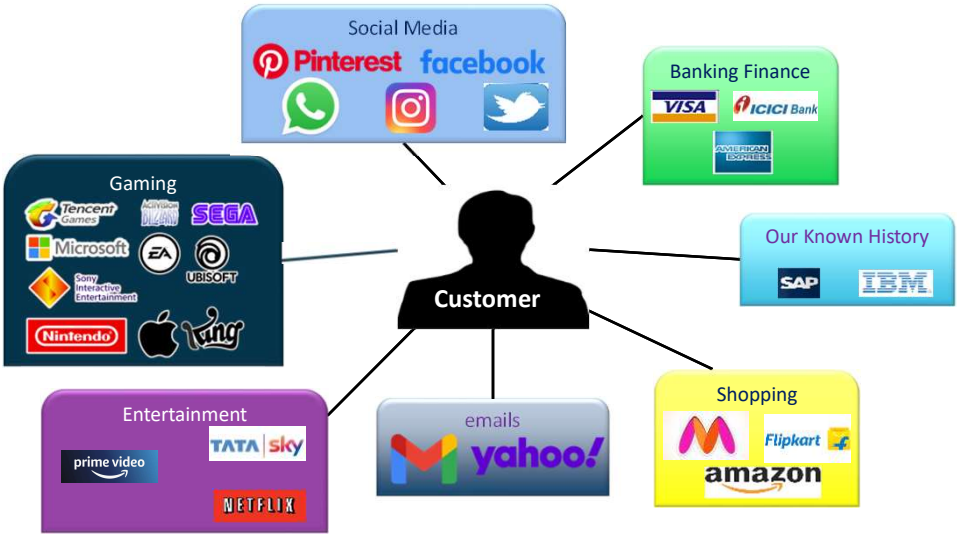
21

Data and AI are Merging into a Synergistic Relationship

- Big data is most assuredly here to stay
  - AI will remain in high demand for the foreseeable future
- AI is useless without data, and mastering data is insurmountable without AI
- Their combination reveals and predicts upcoming trends in business, technology, commerce, entertainment, and everything in between

22

A Single View of the Customer



pra-sâmi

23

Everything You Want to Know About Customer

- Internet now provides a level of concrete information about consumer
- Their habits, likes and dislikes, activities, and personal preferences
  - It was impossible couple of decade ago
- AI's greatest assets is:
  - Its learning ability
  - Capacity to recognize data trends
  - Adaptability to changes and fluctuations
  - Identification of outliers
    - Significant pieces of customer feedback
- AI's ability exploit data analytics
  - AI and Big Data go hand in hand.

pra-sâmi

24 What's Big Data?

No single definition; here is from Wikipedia:

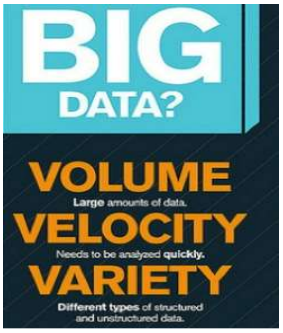
- Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications
- The challenges include
  - capture, curate, storage, search, sharing, transfer, analysis, and visualization
- Enterprise Datawarehouse are preferred over subject-wise datamarts:
  - Allows correlations to be found to "spot business trends", "determine quality of research", "prevent diseases", "link legal citations", "combat crime", "determine real-time roadway traffic conditions", etc.

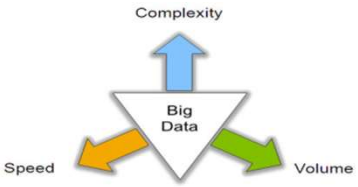
25 Big Data In Simpler Terms

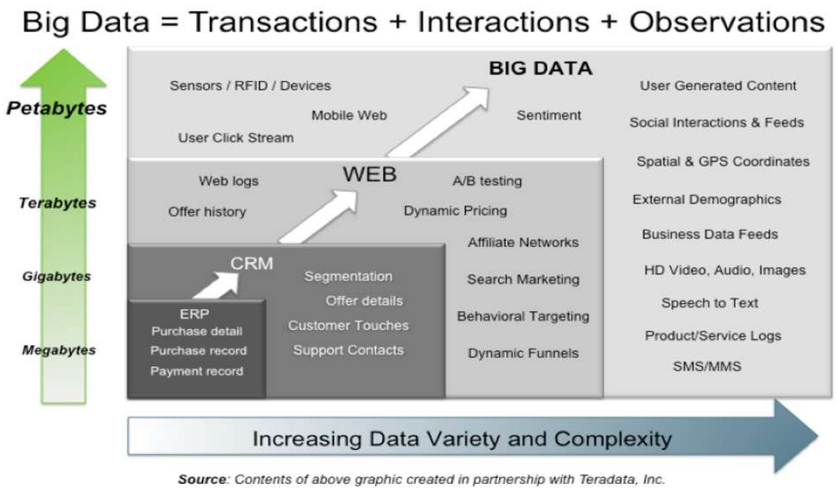
- All data that is not a fit for a traditional RDBMS, whether used for OLTP or Analytics purposes
- So, it requires different approaches:
  - Techniques, tools and architecture
- An aim to solve new problems or old problems in a better way
- Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

26

Big Data: 3V's







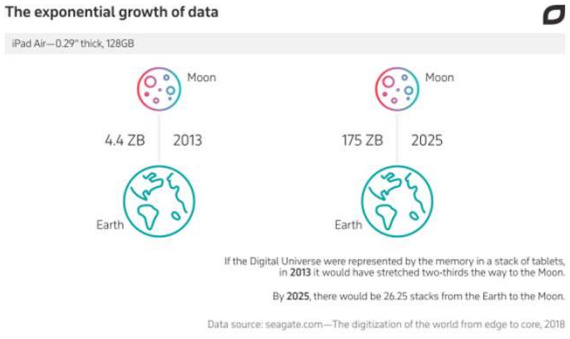
***Source**: Contents of above graphic created in partnership with Teradata, Inc.*

pra-sâmi

27

Volume (Scale)

- Data volume is increasing exponentially
- In its Data Age 2025 report for Seagate, IDC forecasts the global datasphere will reach 175 zettabytes by 2025
- To help you understand how big it is, let's measure this amount in 128GB iPads.
- In 2013, the stack would have stretched two-thirds of the distance from the Earth to the Moon.
- By 2024, anticipated storage is in brontobytes (1 E +27 bytes).
  - Trillions of petabyte
- Walmart is in the process of creating the world's biggest private cloud for processing 2.5 PB of data every hour
- Facebook handles 40 billion photos from its user base.



The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds containing environmental, location, and other information, including video

pra-sâmi

29

CERN’s Large Hadron Collider (LHC) generates 15 PB a year

pra-sâmi

30

The Earthscope

- The Earthscope is the world's largest science project
- Designed to track North America's geological evolution, this observatory records data over 3.8 million square miles, amassing 67 terabytes of data
- It analyzes seismic slips in the San Andreas fault, sure, but also the plume of magma underneath Yellowstone and much, much more

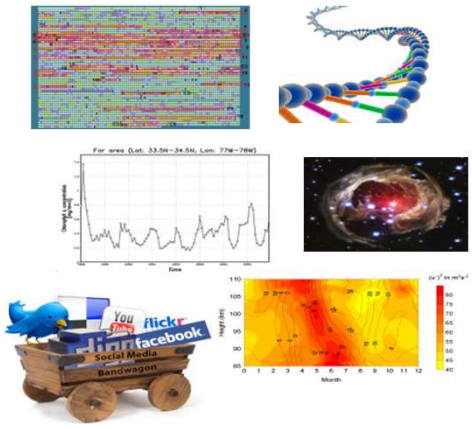
http://www.msnbc.msn.com/id/44363598/ns/technology_and_science-future_of_technology/#.TmetOdQ--uI

pra-sâmi

31 Variety

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML, JSON)
- Graph Data - Social Network, Semantic Web (RDF), …
- Streaming Data - You can only scan the data once
- Big Public Data (online, weather, finance, etc.)
- IOT Data

- Extract knowledge ➔ all these types of data need to linked together



pra-sâmi

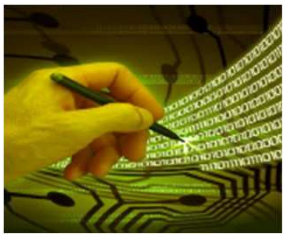
32 Variety

- Big Data isn't just numbers, dates, and strings. Big Data is also geospatial data, 3D data, audio and video, and unstructured text, including log files and social media.
- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.
- Big Data analysis includes different types of data

pra-sâmi

33 Velocity (Speed)

- Data is being generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities



- Examples
  - E-promotions: based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - Healthcare monitoring: sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction
  - Clickstreams and ad impressions capture user behavior at millions of events per second
  - High-frequency stock trading algorithms reflect market changes within microseconds
  - Machine to machine processes exchange data between billions of devices
  - Infrastructure and sensors generate massive log data in real-time
  - On-line gaming systems support millions of concurrent users, each producing multiple inputs per second

34 Real-time/Fast Data

- The progress and innovation is no longer hindered by the ability to collect data
- But, by the ability to manage, analyze, summarize, visualize, and discover knowledge from the collected data in a timely manner and in a scalable fashion

Social media and networks
(all of us are generating data)

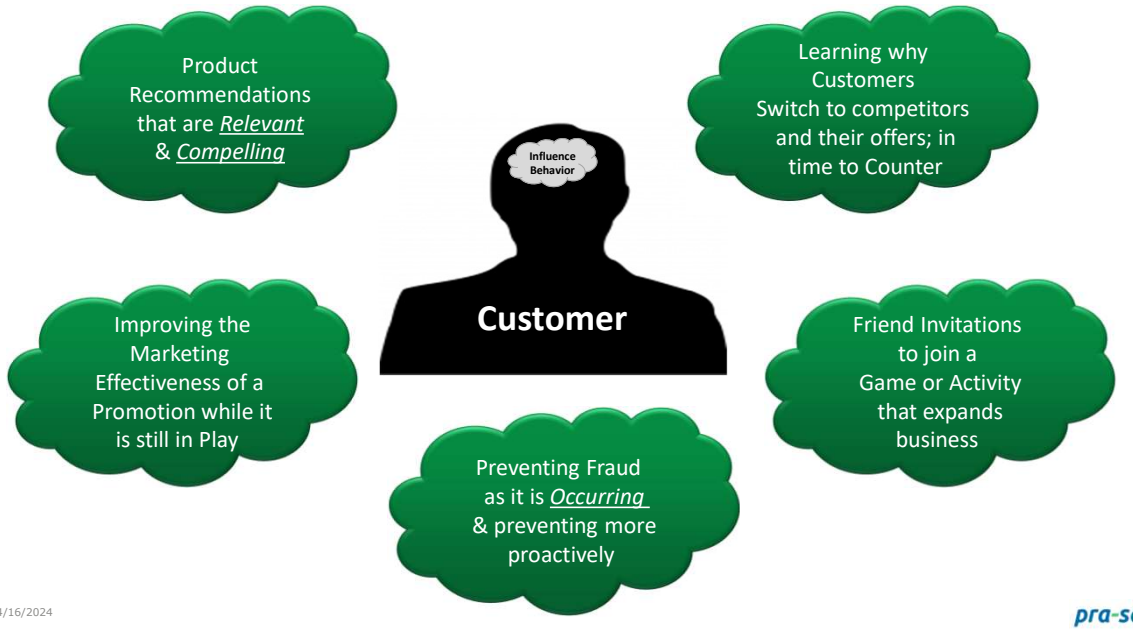
Scientific instruments
(collecting all sorts of data)

Mobile devices
(tracking all objects all the time)

Sensor technology and networks
(measuring all kinds of data)

35

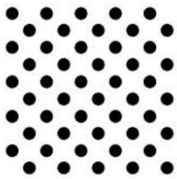
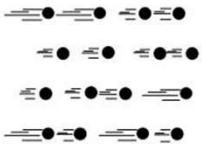
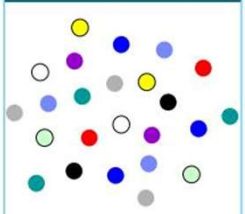
Real-Time Analytics/Decision Requirement



pra-sâmi

36

Some Make it 5V's

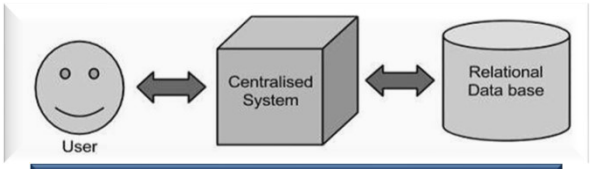
Volume	Velocity	Variety	Veracity*	Value*
				
Data at Rest	**Data in Motion**	**Data in Many Forms**	**Data in Doubt**	**Data of Value**
Terabytes to exabytes of existing data to process	Steaming data, milliseconds to respond	Structured, unstructured, text, multimedia	Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations.	Not useful till it's value is discovered, transform a stream of data into business.

- Later Two more V's were added : Value and Veracity
- Data has no use till it's value is discovered
- Its true and accurate

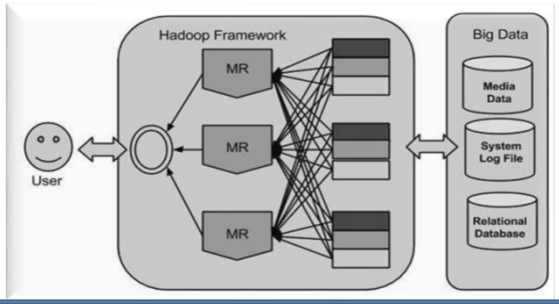
pra-sâmi

37

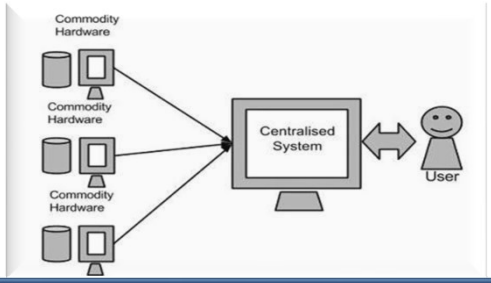
HADOOP



Structured Data –> Maintained by RDBMS software



BIG Data –> Handled by HADOOP (MAP REDUCE FRAMEWORK)

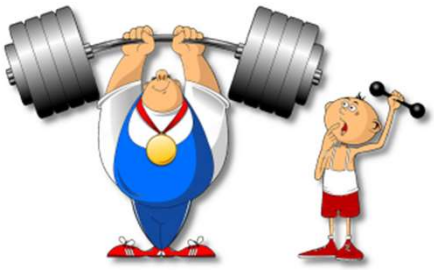


Unstructured Data –> Managed by NOSQL databases

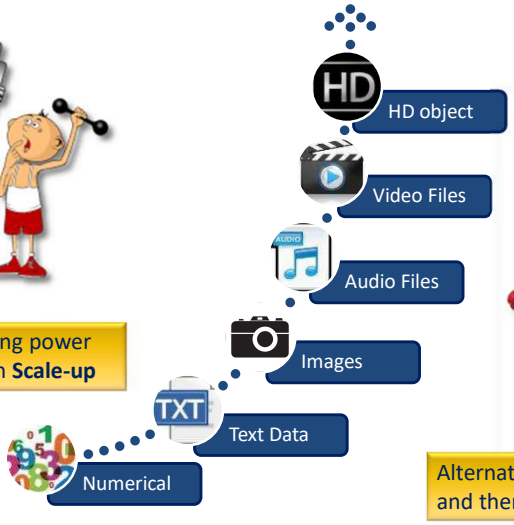
pra-sâmi

38

Scale Up or Scale Out



We can keep adding the processing power but there is a limit to which it can **Scale-up**



Alternatively, we can distribute the work and then compile the results → **Scale-Out**

pra-sâmi

39

Welcome Hadoop

40

HADOOP

- Google solved **Scaling** problem using an algorithm called MapReduce
- This algorithm divides the task into small parts and assigns those parts to many computers connected over the network, and collects the results to form the final result dataset
- Doug Cutting, Mike Cafarella and team took the solution provided by Google and started an Open Source Project called HADOOP in 2005
  - Doug named it after his son's toy elephant
- Hadoop is an Apache open source framework written in java that allows distributed processing of large datasets across clusters of computers using simple programming models
- Hadoop is designed to scale up from single server to thousands of machines, each offering local computation and storage

41

HADOOP

- Hadoop is not a type of database, but rather a software ecosystem that allows for massively parallel computing
- It is an enabler of certain types NoSQL distributed databases (such as HBase), which can allow for data to be spread across thousands of servers with little reduction in performance.
- Hadoop provides two things:
  - A distributed filesystem called HDFS (Hadoop Distributed File System)
  - A framework and API for building and running MapReduce jobs

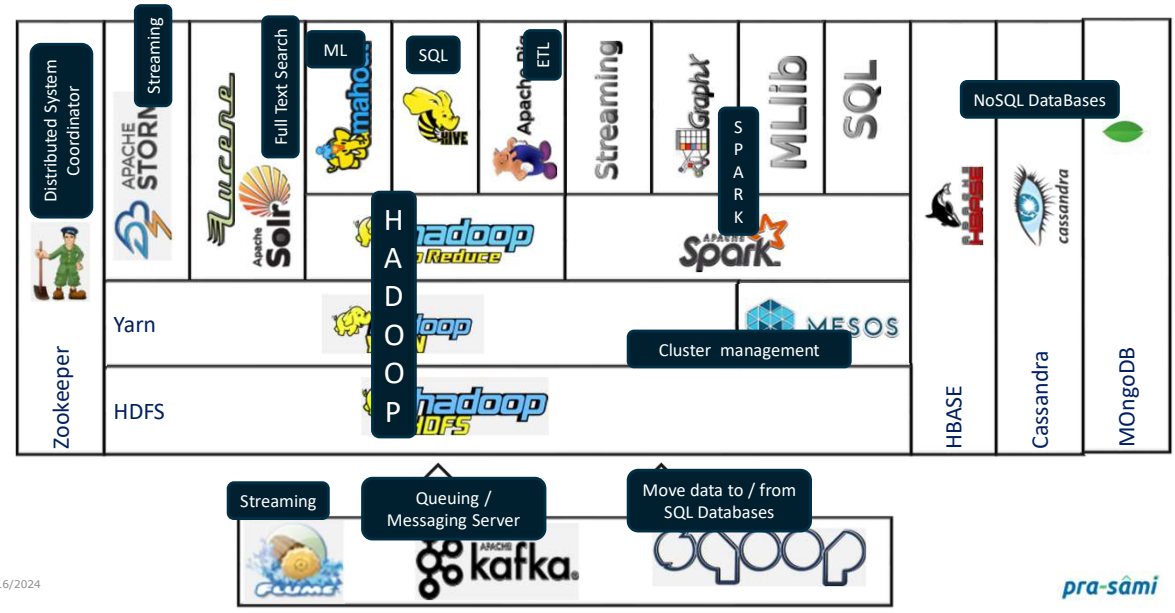
42

Hadoop

- Ability to scale out to Petabytes in size using commodity hardware
- Processing (MapReduce) jobs are sent to the data versus shipping the data to be processed
- Hadoop doesn't impose a single data format so it can easily handle structure, semi-structure and unstructured data
- Hadoop is a giant storage pool with mostly batch oriented ways to retrieve the large datasets and apply distributed computing methods
- HDFS is structured similarly to a regular Unix file system except that data storage is distributed across several machines.

43

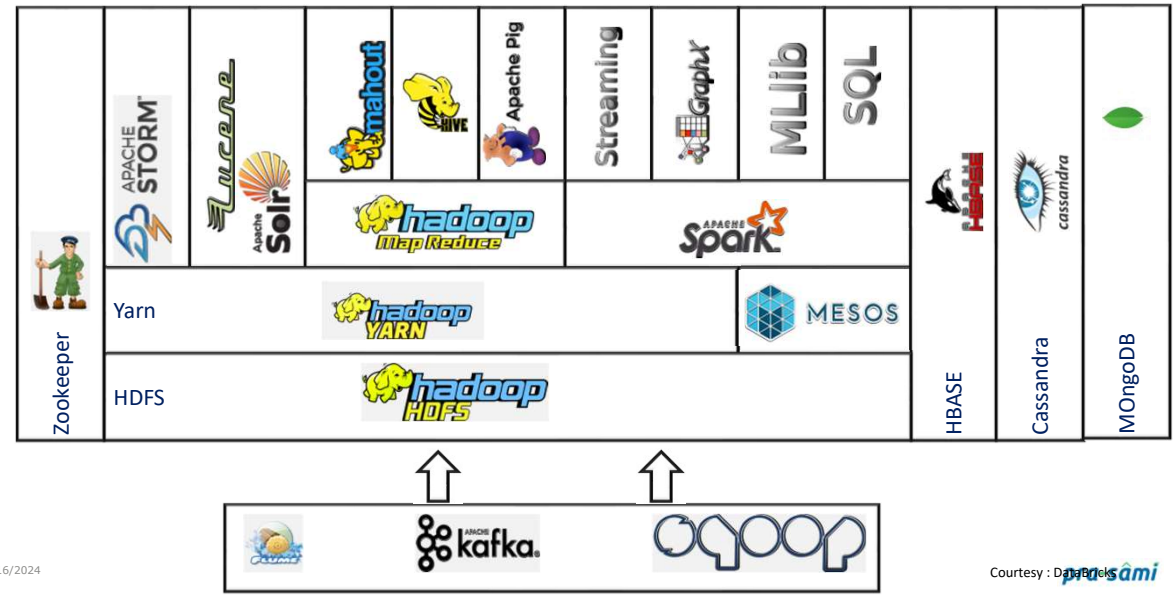
Hadoop Ecosystem



pra-sâmi

44

Hadoop Ecosystem



Courtesy : DataBricks

pra-sâmi

45

AI Using Big Data

- Natural language processing:
  - Millions of samples from the human language are recorded and linked to their corresponding computer programming language translations
  - Models are prepare and used in helping organizations analyze and process huge amounts of human language data.
- Helping agricultural organizations and corporations :
  - Broaden their monitoring capability, including Geo tagging of the products
  - AI helps farmers to count and monitor their produce through every growth stage till maturity
  - AI can identify weak points or defects long before they spread to other areas of these huge acres of land
  - In this case, satellite systems or drones are used by the AI for viewing and extracting the data.
- Banking and securities:
  - Monitoring financial market activities. For instance, the Securities Exchange Commission (SEC) is using network analytics and natural language processing to foil illegal trading activities in financial markets
  - Trading data analytics are obtained for high-frequency trading, making decision-based trading, risk analysis, and predictive analysis
  - They are also used for early fraud warning, card fraud detection, archival and analysis of audit trails, reporting enterprise credit, customer data transformation, etc.

46

AI Using Big Data

- Communication, Media and Entertainment:
  - AI capabilities can be used for collecting, analyzing, and utilizing consumer insights
  - Leveraging mobile and social media content
  - Understanding patterns of real-time, media content usage
  - Companies can simultaneously analyze their customer data along with customer behavioral data to create detailed customer profiles that will be used for creating content for a diverse target audience, recommending content, and measuring content performance.
- Healthcare provider:
  - Have benefited from the large pool of health data Prescriptions and health analysis have been simplified by AI
  - Hospitals are using data collected by millions of cell phones and sensors, allowing doctors to use evidence-based medicine
  - Also, the spread of chronic diseases is identified and tracked faster.
- Educational sector:
  - AI syncs with Big Data analytics for various purposes, such as for tracking and analyzing when a student logs into the school's system, the amount of time spent on the different pages of the system, and the overall progress of students over time
  - It is also useful for measuring the effectiveness of teachers
  - Thus, teachers' performance is analyzed and measured with respect to the number of students, various courses, student aspirations, student demographics, behavioral patterns, and many other data

47

AI Using Big Data

- Manufacturing, inventory management, production management, supply chain analysis and customer satisfaction techniques are made seamless
  - Quality of products is improved, energy efficiency is ensured, reliability levels rise, and profit margins increase.
- Natural Resources sector:
  - The synergy of AI and Big Data makes predictive modeling possible
  - Allowing for the quick and easy analysis of large graphical data, geospatial data, temporal data, seismic interpretation and reservoir characterization
- Governments around the world use AI for various applications such as public facial recognition, vehicle recognition for traffic management, population demographics, financial classifications, energy exploration, environmental conservation, infrastructure management, criminal investigations, and much more
- Other areas where AI is used in Big Data are Insurance, Retail & Wholesale Trade, Transportation, and Energy & Utilities

48

Agenda

49