Big Data

Big Data - is data that exceeds the processing capacity of conventional database systems. Collection of data sets so large and complex that it becomes difficult to process using on-hand data management tools.

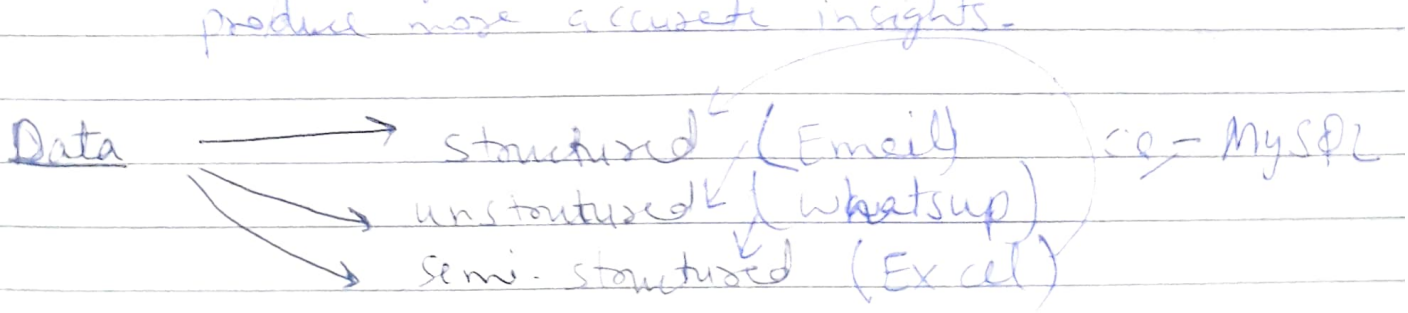
Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.

Where Is This 'Big Data' coming From?
social data, machine data, transactional data.

Big Analytics

process of examining data - typically of a variety of source types, volumes and/or complexities - to uncover hidden pattern, unknown correlations and other useful information
⇒ intent is to find bussiness insights.

- Deeper insights - Rather then looking at classifications groups, you will insights into all the individual.
- Broader insights - take account in all the data to produce more accurate insights.



Structured data - data stored in rows and columns, mostly numerical, where meaning of each data item is defined. It consists of 10% of total data and is accessible through database management systems or hadoop

DOMS | Page No.
Date / /

Unstructured data - data of different forms like eg. text, image, video, document etc. It can also be in the form of customer complaints, contracts, or internal emails. It comprises of 90% of data.
Unstructured data cannot be stored using traditional relational databases. Storing data with such a variety and complexity requires the use of adequate storage systems, commonly referred to as NoSQL database eg - MongoDB

- semi-structure data and unstructured data can only be handle through Hadoop system.
- CRUD CYCLE is not supported by Hadoop.

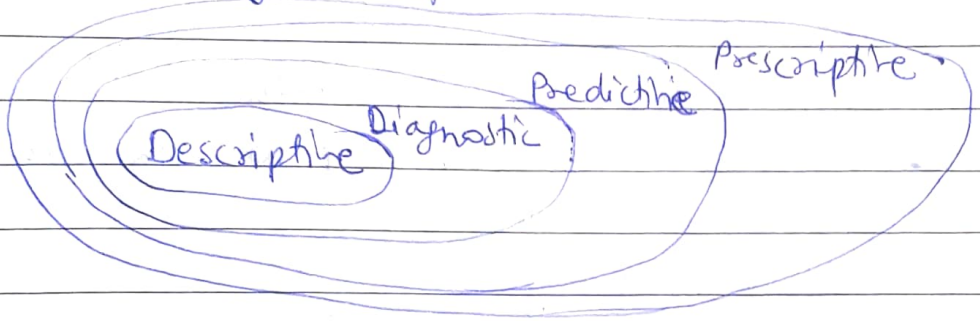
Semi-Structured data - is a form of structured data that does not obey the tabular structure of data model associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to seperate semantic elements and enforce hierarchies of records and fields with data.

What happen and why?

Diagnostic data - is data that is automatically recorded by S/w for the purposes of troubleshooting problems.
form of advance analysics which examines data or content to answer the qu
Descriptive data - data describe the basic features of the data in a study and provide summery about the samp and the measures.

Predictive Analytics - is the use of data, statical algorith and machine learning techniques to identify the likelihood of future outcomes based on historical data.

Descriptive - analytics addresses the issue of what happened.
Diagnostic - analytics answers the question of why something happened.
Predictive - analytics describes what is likely to happe
Prescriptive - analytics prescribes what steps to take to avoid a future problem



⇒ visualization tools (weka Tool)

	Structured Data	Semi Structured Data	Unstructured Data
nology ty rmance	Relational Database Table Structured query allows complex joins	XML/RDF Queries over anonymous nodes are possible	character/binary dat Only textual queries possible
lability	scaling DB schema is difficult	simple	very scalable

3 Dimensions / characteristic of Big data
3V are three defining properties or dimensions of big data. Volume refers to amt. of data, variety refers to the number of types of data and velocity refers to the speed of data processing.

Volume - Size of available data has been growing at an increasing rate.
Velocity - Data is increasingly accelerating the velocity at which it is created and at which it is integrated.

Date / /

Variety - presents an equally difficult challenge. The growth in data sources has fuelled the growth in data types. In fact, 80% of world's data is unstructured

veracity - refers to the biases, noise and abnormality in data Is the data that is stored meaningful to the problem being analyzed. (inconsistencies and uncertainty in data)

Value - Bulk of Data having no value is of no good to the company, until you turn it into something useful.

Problem with Big Data

i) Lack of Understanding ii) Security Gaps
iii) Low Quality and Inaccurate Data iv) Lack of skilled

Dataset - is a collection of data. In case of tabular data a dataset corresponds to one or more database tables, where every column of a table represents a particular variable and each row corresponds to a given record of the data in question.

DOMS | Page No.
Date / /

Case Study

Manufacturing Big Data Use Cases

i) Predictive Maintenance

- helps in predict equipment failure.
- deploy maintenance more cost effectively. data can be get from sensor data, engine temperature
- to predict the remaining optimal life of systems and components to ensure that they perform within specifications.

challenges

- integrate data coming from different formats and identify the signals that will lead to optimizing maintenance.

ii) Operational Efficiency

- analyze and estimate production processes, proactively respond to customer feedback, and anticipate future demands.

challenges

Data teams must balance the data volume with the growing number of sources, users, and application

iii) Production Optimization

- understanding the flow of goods in production
- Using Data analysis to increase productive time.

challenges

Combining different kinds of data.

DOMS | Page No.
Date / /

b) Retail Big Data Use Cases

a) Product Development –

By classifying key attributes of past and current products and then modeling the relationship b/w the attributes, building predictive models for new produ and services. Deeper analysis to plan, produce and launch new products.

Challenges

- Creating segments acc. to customer behaviour identify sophisticated use patterns and behaviour and map them to potential new offerings.

b) Customer Experience

- gathering data from social media, web visits, call lo and other data source. Big data analysis can be used to deliver personalized offers, reduce customer churn, and proactively handle issues.

Challenges high vol

- gathering data from various resource can be difficult. Once the data is integrated, path analy can be used to identify experience paths and correlate them with various set of behaviour

c) Customer Life Time Value

- identify your best customers, marketing them with special offers. Sales team devotes more time to them.

Challenge

- Identify your high-value customers, analysing a high vol. of transaction data and create sophisticated model that examine past behaviour and predict future actions.

DOMS | Page No.
Date / /

Healthcare Big Data Use Cases

a) Genomic Research

Using big data, researchers can identify disease genes and biomarkers to help patients pinpoint health issues they may face in the future.

challenges

- running complex algo. on the data is complicated and can require long processing times

b) Patient Experience and outcomes

With big data, health care organization can creat 360° view of patient care as patient moves through various treatments and departments.

challenge

- Improving the patient experience requires a large volume of patient data (it can be multi-structure)

c) Claims Fraud

Big data helps healthcare organizations detect potential fraud by flagging certain behaviours for future examination.

challenge

- Claims fraud analytics is a complex process that involves integrating different datasets, analyzing the claims data, and identifying complex fraud patterns.

Oil and Gas Big Data Use Case

a) Predictive Equipment maintenance

- predict the remaining optimal life of their systems and components, ensuring that their assets operate at optimum productive efficiency.

challenge

collecting data of varying format.

b) Oil Exploration and Discovery

Data generated from seismic monitors can be used to find new oil and gas sources by identifying traces that were previously overlooked.

Challenges

- integrate and analyze an enormous volume of structure data

c) Oil Production Optimization

- measures well production to understand usage rates. Analyse why well outputs aren't tally with their predictions.

Challenges

- analyzing large volume of data

e) Telecommunications Big Data Use Cases

a) Optimize Network capacity

Network usage analysis can help companies identify areas with excess capacity and reroute bandwidth as need, BD helps in planning for infrastructure investment and design new services that meet customer demand

Challenges

for network analytics requires analyzing a high volume of call detail records

b) Telecom customer churn

- predict overall customer satisfication by considering service quality, convenience
- set up alert when customers are at the risk of churning - and take action with retention campaign and proactive offers

Challenges

- require to analyzing past and current data to create new model to predict churn.

c) New product Offerings
- design new products and features based on customer behaviour. (for future offerings)

Challenge
analyzing high-volume product-log data in different format.

Crowdsourcing is the practice of engaging a 'crowd' or group for a common goal - often innovation, problem solving or efficiency. It is powered by new technologies, social media. Crowdsourcing can take place on many different levels and across various industries.
- collectively contribute - whether with ideas, time, expertise or funds - to a project or cause.
- powerful bussiness marketing tool - promoting and growing the company by creativity.

Advantages
i) Numerous idea from numerous people
ii) cheap - similiar to outsourcing, It is used to cut costs, when you don't have to employ people and pay them a wage, whether they are working or not, it definitely lowers costs
iii) fast - take less time to find right person to do the job

Disadvantage
i) Quality could be questionable - when you hire numerous people to do a job, if you aren't lucky enough to hire the professionals.
ii) unreliable way to get a job done
iii) confidentially

DOMS | Page No.
Date / /

Inter Firewall & bans Firewall Analysis

	Cloud Computing	Big Data
Definition	• Provides resources (storage, computing dB, monitoring tools etc) on demand.	• Provide a way to handle huge volumes of data and generate insights.
Reference	• It refers to internet services from SaaS, PaaS to IaaS.	• It refers to data, which can be structured, semi-structured and unstructured.
Use for	• Use to store data and info. on remote servers (range of network of cloud servers over the internet)	• used to describe huge vol. of data and info (discover undiscovered patt

Cloud Computing Services

i) Infrastructure as a Services (IAAS) - Here the service provider offers entire infrastructure along with the maintenance related tasks. Organization make use of unlimited storage potential of the cloud infrastructure. They can expand and shrink their storage space as needed w/o worry.

ii) Platform as a Service (PAAS) - In this service, the cloud provider offers resources like queuing, databases etc. However, the responsibility of configuration and implementation related tasks depends on the customer.

iii) Software as a Service (SAAS) - This service is the most facilitated one which provides all the necessary setting and infrastructure provides IaaS for the platform and

infrastructure are in place. (provides self service capabilities to users with scalable features to increase usage on requirement)

Cloud computing types

i) private - used for a single organization, can be internally or externally hosted.

ii) public - Provisioned for open use for the public by a particular organization, who also hosts the service.

iii) Hybrid - composition of 2 or more clouds (private, public or community), that remain unique entities but they bound together.

iv) community - shared by several organization, typically externally hosted.

G2G, G2B, G2C, G2E

R-language
SPSS S/W
IBM watson
] virtual Tool.

Mid sem

RDB
ACID

NoSQL
BASE
Basically, Available, Soft-state, Eventually co

	SQL	NoSQL
Type of db	Relational database	Non-relational Database
schema	Pre-defined	Dynamic
Db categories	Table based dB	Document-based dB, key-value stores, Gra
Complex queries	Good for CQ	Not good for CQ
Hierarchical Data storage	Not the best fit	best fit
scalability	Vertically scalable	Horizontally scalable
Languages	Structured Query Language	Unstructured Query lang
Base Properties	Based on ACID property	Based on CAP Theorem

NoSQL- database provides a mechanism for storage and retrieval of data that is modelled in means other than the tabular relations. (Not only SQL)

Theory of NOSQL : CAP

- Many Nodes
- Nodes contain replicas of partitions of the data.

Consistency
→ All replias contain the same version of data
→ client always has the same view of the data (no matter what node)

CAP

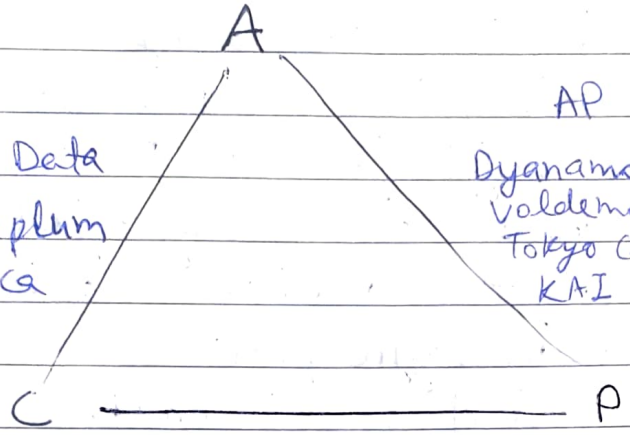
Availability
→ system remains operational on failing nodes.
→ All clients can always read and write.

Partition tolerance
→ multiple entry points
→ System remains operational on system split (communication malfunction)
→ System works well across physical network partitions.

Types of NoSQL

DOMS | Page No.
Date / /

Use of NoSQL



CA
BMS
SQL,
stores,
etc)
Aster Data
Greenplum
Vertica

AP
Dyanamo
Voldemort
Tokyo Cabinet
KAI
Cassandra
SimpleDB
CouchDB
RIAK

CP
Big Table
Hypertable
Hbase
MongoDB
Terrastore
BerkeleyDB
MemcacheDB

Characteristic / features of NOSQL

- Non-relational
- open-source
- schema-free
- simple API
- Build for new generation web applications
- distributed
- horizontally-scalable
- easy replication support
- high amt. of data

NoSQL is an approach to database design that can accomodate a wide variety of data model.
NoSQL which stands for 'Not only SQL' is an alternate to traditional relational database.
NoSQL is specially designed for large amt. of data stored in distributed environment.

Need - bcz data are not structured eg data at facebook, google.

Advantage - i) supports all type of data
ii) No static scheme
iii) Faster data processing
iv) low operational cost
v) support Distributed system

- Elastic scaling - is the ability to automatically add or remove networking infrastructure based on changing application traffic patterns. Elastic load balancer with scaling is used to automatically adjust the amt of resources that are allocated to deliver an application in response to changes in traffic patterns.
- Big Data - Vol of data that are being stored have increased.
- Goodbye DBA's - Automatic repair, distribution, tuning.
- Economics -
  cheap commonity servers ⇒ less costs per transaction
- Flexible Data models - allows the direct representation of data with irregular scheme.

Disadvantages
i) Limited query capabilities.
ii) Not a defined standard.

Challenges
i) Maturity → still in pre-production phase
   ↳ key features yet to be implemented
ii) Support → limited resources or credibility
iii) Administration
   - Require lot of skill to install and effort to maintain.
iv) Limited ad-hoc querying
v) Expertise - Few no. of NoSQL experts available in the market.

RDBMS	NoSQL
• regular backup	• replication
• access through master server.	• sharding across multiple nodes
• non-programmers writing queries.	• only programmers writing queries.
• data updates are frequent	• write-once & read many times.

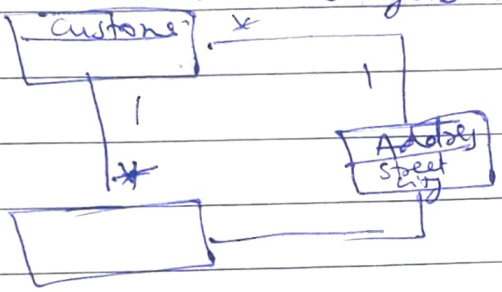
- predictable, linear growth

- ~~only programmers writing queries~~

DOMS | Page No.
Date / /

Data Model

Aggregate - is collection of related ~~objects~~ that we treat as a Unit. When a unit of data is retrieved from the db gets all the related data along with it.



These units of data form the boundary for ACID operation with the 4 dB.

Type of NoSQL database

- Key values
- Document oriented
- Wide Column family
- Graph databases

Key values

Key value

↑ lookup

id	name	age

eg- ~~HBase~~, Redis

Every single item in the dB is stored as an attribute name (or "key") together with its value stores

Document oriented

eg-

id	age
1	10
2	20

```
{ "id" : 1
  "name" : "A"
  "age" : 10
}
```

document & collection (table)

eg - MongoDB

Documents can contain many different key-values pairs

Wide Column family

Columns collect ↓ column family

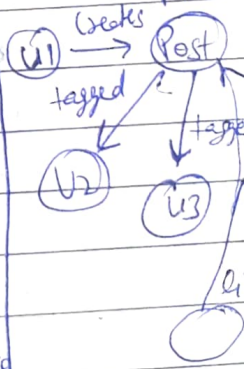
key	personal		profession	
emp_id	→ name	→ age	→ exp	→ sal
	→ loc			

eg - Cassandra, Hbase

optimized for queries over large datasets and stores cols of data together, instead of rows

Graph databases

eg - Neo4j, TitanDB



relationships

are used to store info about networ

DOMS | Page No.
Date / /

Big Table – is a distributed Storage system for manag structured data at Google.
↳ scalability, high performance
↳ Sparse, distributed, multidimensional sorted m

API – • functions for creating and deleting tables & column families etc.

Master Server – responsible for assigning tablets to tablet servers and balancing the tablet server load, garbage collect

tablet – set of consecutive rows of a table and is the unit of distribution.

Tablet server – • Each tablet server manages a set of tab
• clients communicate directly with the tablet serve
• Tablet server handles read and write requests
• Also split tablets that has grown too large

Materialized view – is a database object that contains the results of a query.

DOMS | Page No.
Date / /

MongoDB → CP

It is a free and open-source cross platform document-oriented database program.

It uses JSON-like documents with schemes.

Uses BSON format

It is a binary form for representing simple data structures

BSON ⇒ "Binary JSON"

Which PL can be used with MongoDB?

JS, Python, Ruby, Perl, Java, C#, C++.

Functionality of MongoDB

- Dynamic schema
- Document-based database
- Built-in horizontal scaling
- secondary indexes
- Query language via an API
- Querying
- Fast In-place Updates

Why use MongoDB?

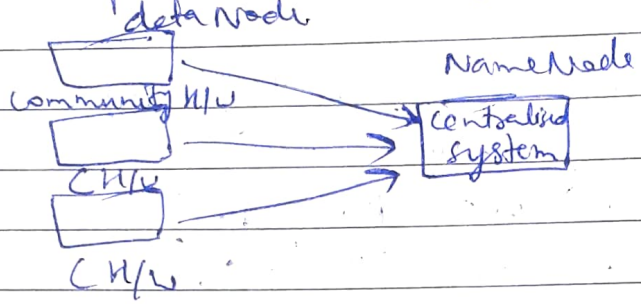
- simple queries
- Easy & fast integration of data
- No ER diagram
- Rich queries
- Document Oriented storage

RDBMS	MongoDB
Table, view	collection
Row	Document [BSON]
col	Field
Join	Embedded Document

DOMS | Page No.
Date / /

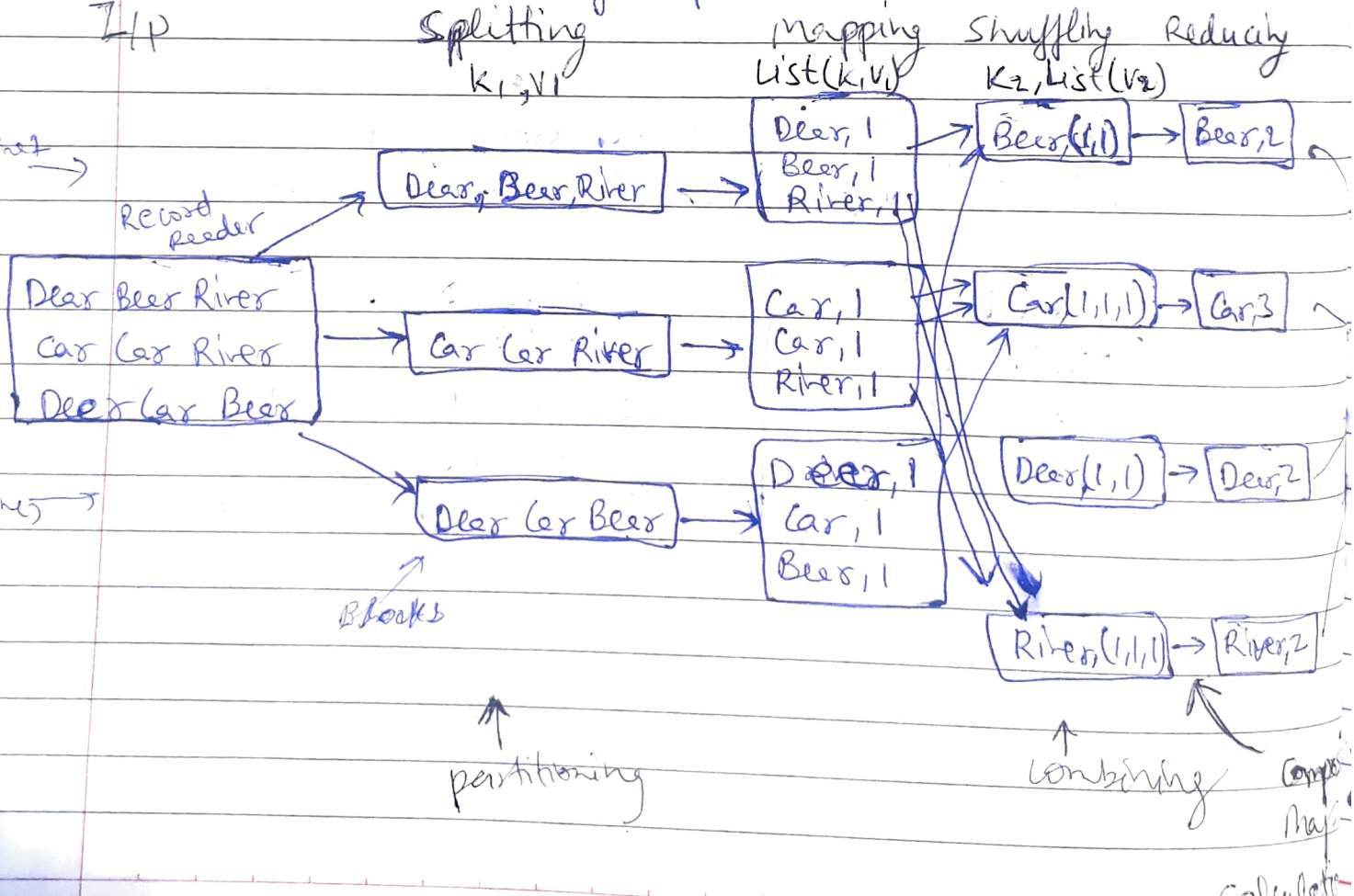
Map Reduce

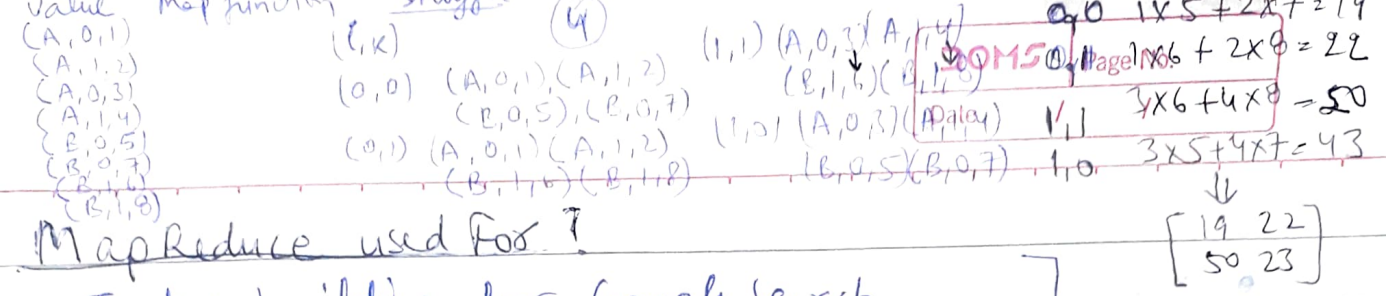
(simple programming model for data-intensive computing on large commodity clusters
Map Reduce divides a task into small parts and assig
them to many computers. Later, the results are collecte
at one place and integrated to form the result datas



- The Map task takes a set of data and converts it into another set of data, where individual elements are broken down into tuples (key-value pairs)
- The Reduce task takes the o/p from the Map as an input and combines those data tuples (key-value pairs) into a smaller set of tuples.

Working





MapReduce used For ?

- Index building for Google search
- Analyzing search logs (Google Trends)
- Statistical language translation

Google

- Data mining
- Ad optimization
- Machine learning

FB

MapReduce Goals

- Scalability to large data volumes.
- Cost - efficiency.

Commodity cluster computing - involves the use of large numbers of already-available computing components for parallel computing, to get the greater amount of useful computation at low cost.

↳ challenge

- Programming in a cluster is hard.
- cheap nodes may fail
- commodity network → low BW

(V3)

Application

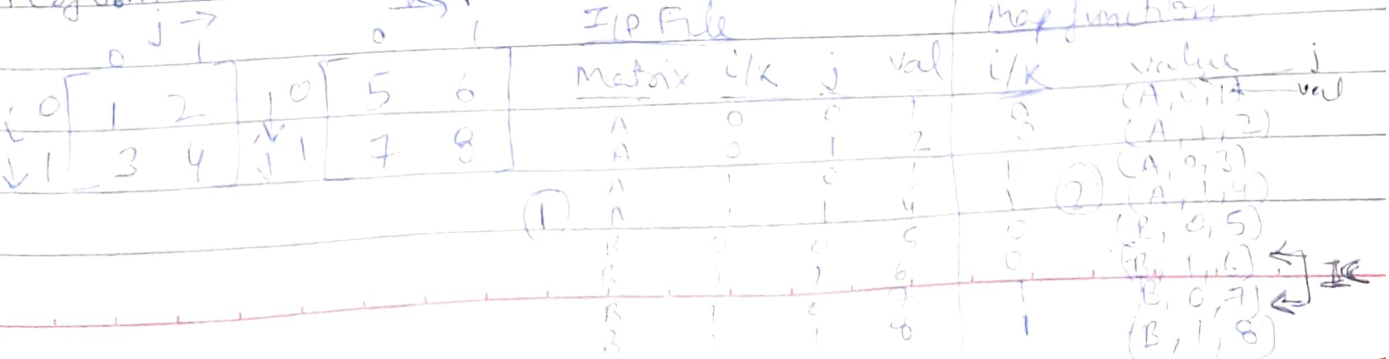
- Google Translator
- Geographical Data
- ~~Ai~~ AI
- Pdf Generator

Matrix Multiplication

1 step 2 step

i) Map
ii) Reduce

Perform matrix multiplication (one step on following matrices)



Cluster - group of system.

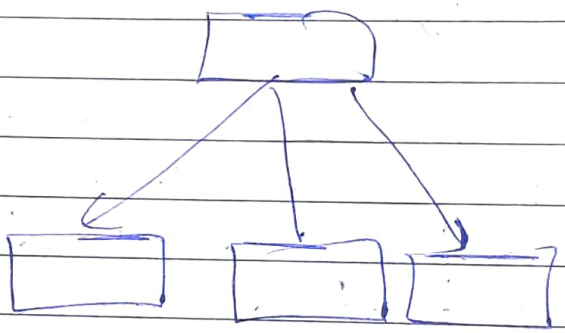
DOMS | Page No.
Date / /

Hadoop

- Open source framework that allows distributed proces of large datasets on the cluster of commodity hardware
- Hadoop is a data management tool and uses Scale out storage

↓

is network-attached storage (NAS) architecture in which the total amount of disk space can be expanded through the addition of devices in connected arrays with their own resources. (H/w added as per need arises)



Hadoop components.

- Hadoop1 components
  - HDFS (Hadoop Distributed File system)
    - ↳ It is storage unit of Hadoop.
  - MapReduce:
    - ↳ It is the processing unit of Hadoop.
- Hadoop 2 components
  - HDFS
  - YARN / MRV2
    - ↳ resource management unit of Hadoop

5 Hadoop Daemons

i) NameNode - Used to hold the metadata (info. about location, size of file/block). There will always be only 1 NameNode in a cluster.

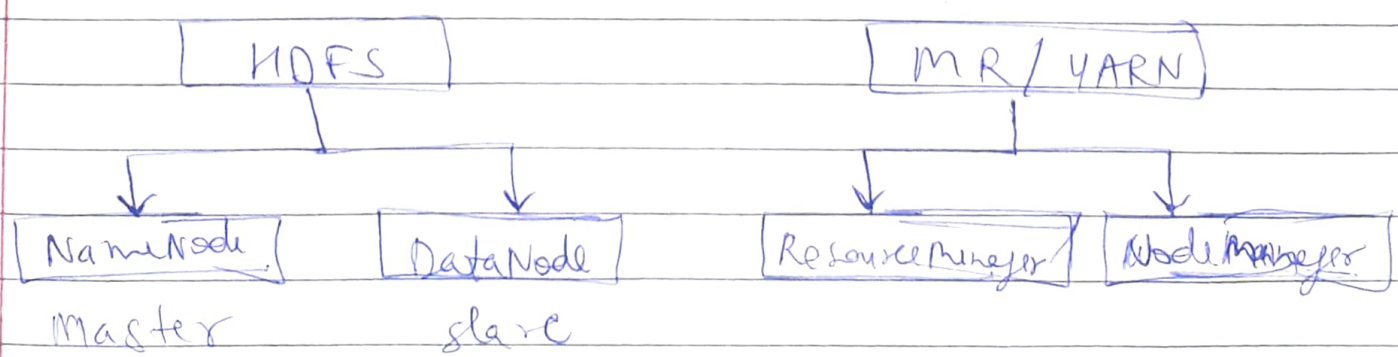
ii) Secondary NameNode - Used as a backup for NameNode. It holds pretty much same information as that of NameNode. If NameNode fails, this one comes into picture.

iii) DataNode - Stores actual HDFS data blocks. The no. of DataNode depends on your data size. The DataNode communicates to NameNode on frequent basis.

iv) JobTracker - Manages MapReduce jobs, distributes individual tasks to machines running the TaskTracker. Coordinates MapReduce stages.

v) TaskTracker - The jobs given by JobTrackers are actually performed by TaskTrackers. Each DataNode will have one task tracker. TaskTrackers communicate with JobTrackers to send statuses of the jobs.

Note - Hadoop HDFS Architecture follows a Master/Slave Architecture, a cluster comprises of a single NameNode (Master Node) and all the other nodes are DataNodes (Slave nodes).



DOMS | Page No.
Date / /

Running Modes of Hadoop

i) Standalone Operation - By default, run in a non-distributed mode (single JVM).
Suitable for running MapReduce programs during Development.

ii) Pseudo - Distributed Mode - Hadoop daemons run on local machine. Each Hadoop daemon will run as a seperate java process.

iii) Fully-Distributed Mode - Hadoop daemons run on a cluster of machines. Fully distributed with minimum 2 or more machines as a cluster.

File Formats in Hadoop

i) Text / CSV Files
ii) JSON File
iii) AVRO Files
iv) Sequence File
v) RC Files
vi) ORC Files
vii) Parquet Files

Hadoop Data Analysis Techniques

to analyze the huge stock data being generated very frequently.

- MapReduce
  - → Powerful model for parallelism
  - → Based on a rigid procedural structure
- Pig
  - → Procedural data-flow language
  - → Used by programmers and researchers
- Hive
  - → Declarative SQLish language
  - → Used by analyst for generating reports.

Scale out - is a type of capacity expansion ~~concentrating~~ on the addition of new N/w resources instead of increasing the capacity of already available H/w resources such as storage.

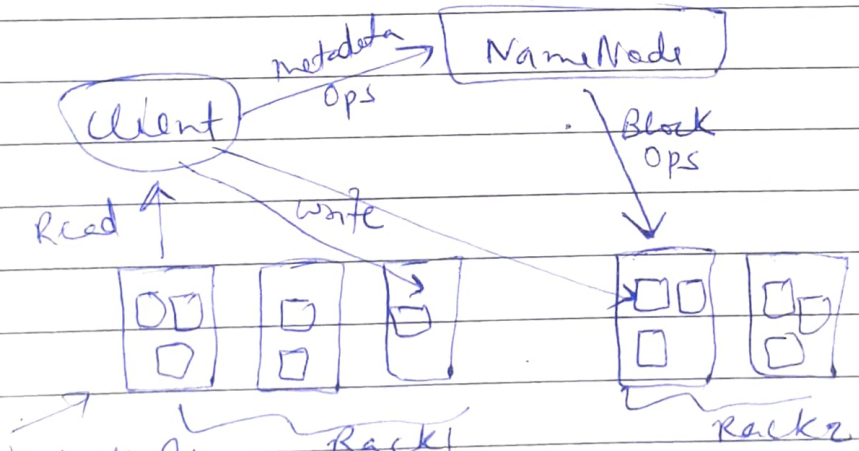
JOMS Page No. Date / /

Hadoop streaming - is a utility that comes with Hadoop distribution. This utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.

Hadoop Pipes - is the C++ interface to Hadoop Reduce. Hadoop Pipes uses sockets to enable task trackers to communicate processes running the C++ map or reduce functions.

HDFS

- scalability for large data set.
- Reliability to cope with H/W failure.
- suitable for the distributed storage & processing
- good for streaming data.



HDFS Architecture

Goal

- Auto Fault detection and recovery
- Huge datasets.
- easy to read

Replication of blocks for fault tolerance

Execution

calling create() by client

↓ RPC call

Name node → check for file already open & permission → NO

↓ yes

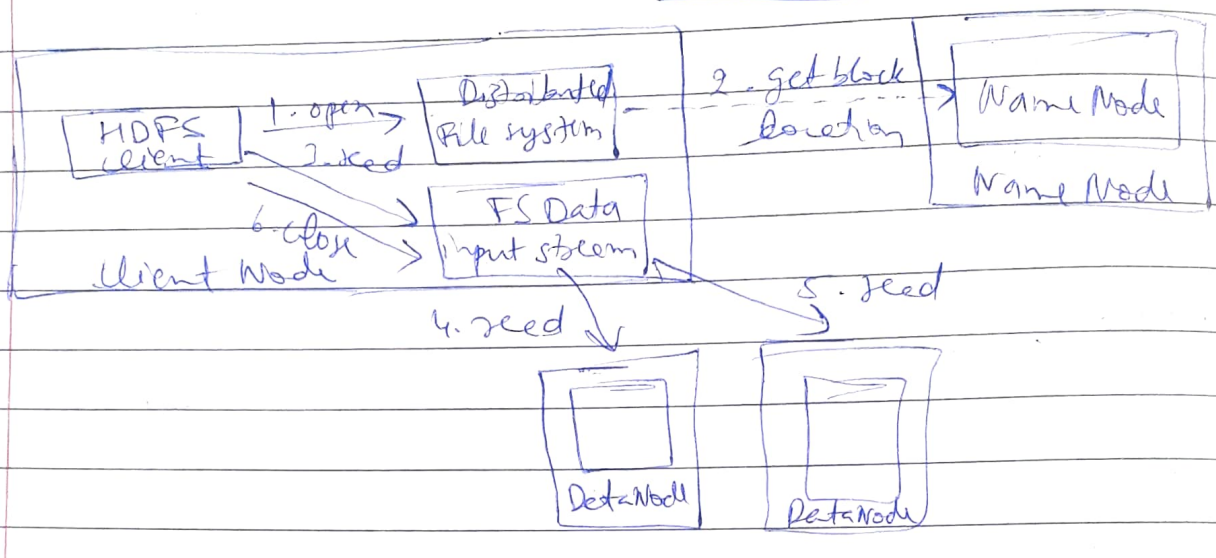
create new file → written by client

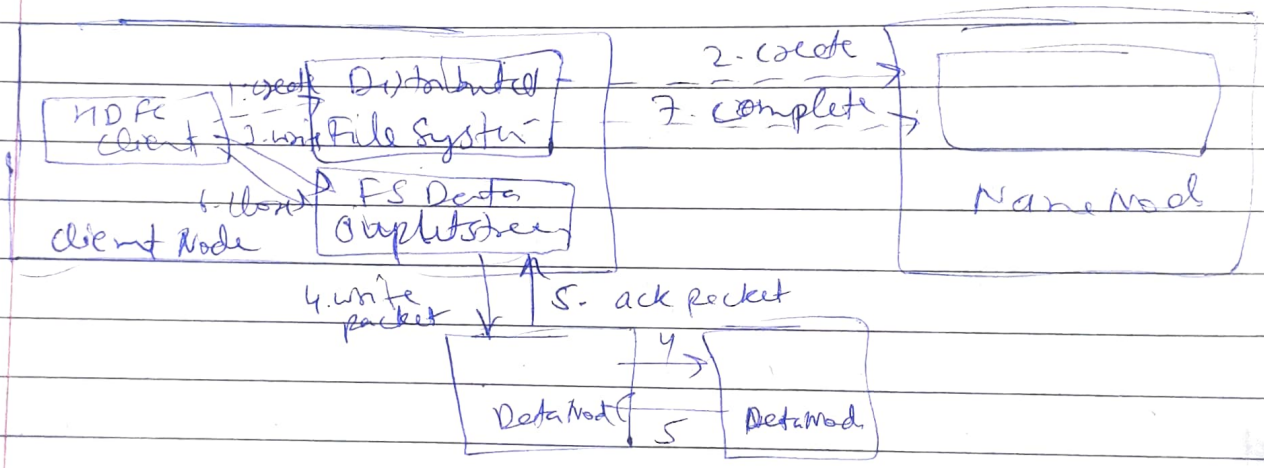
Client - set of programs

DOMS | Page No.
Date / /

I/O in Hadoop

File Read





Data Integrity

- A HDFS client creates the checksum of every block of its file and store it in hidden files in the HDFS namespace
- When a client retrieves the contents of file, it verifies that the corresponding checksums match.
- If does not match, the client can retrieve the block from a replica.

Compression

- Store data in a format that requires less space than original.
- Useful in storing & transmitting the data
- lossless, ~~lossy~~ compression

DOMS | Page No.
Date / /

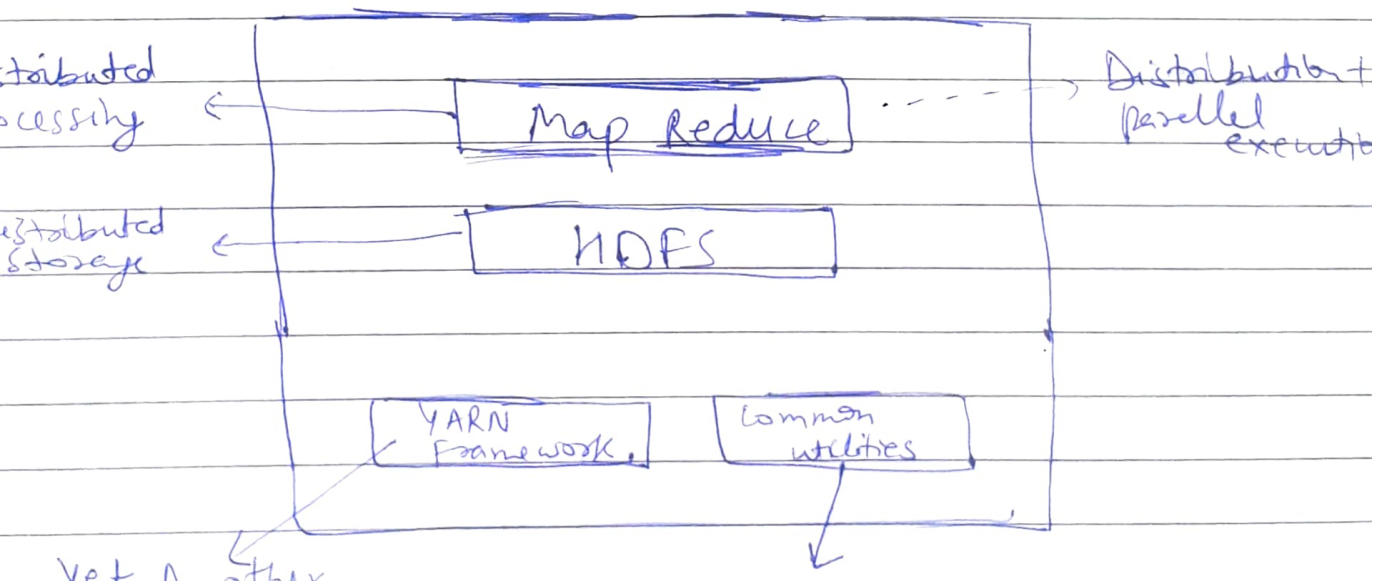
Data serialization

It is a process to format structured data in such away that it can be reconverted back to the original form. It is done to translated data structures into a steam of data.

Avro File-based data structures

Avro is a row-based storage format for Hadoop which is widely used as a serialization platform. Avro stores the data definition (schema) in JSON format makin it easy to read and interpret by any program. The data itself is stored in binary format making it compact & effici

Hadoop Architecture



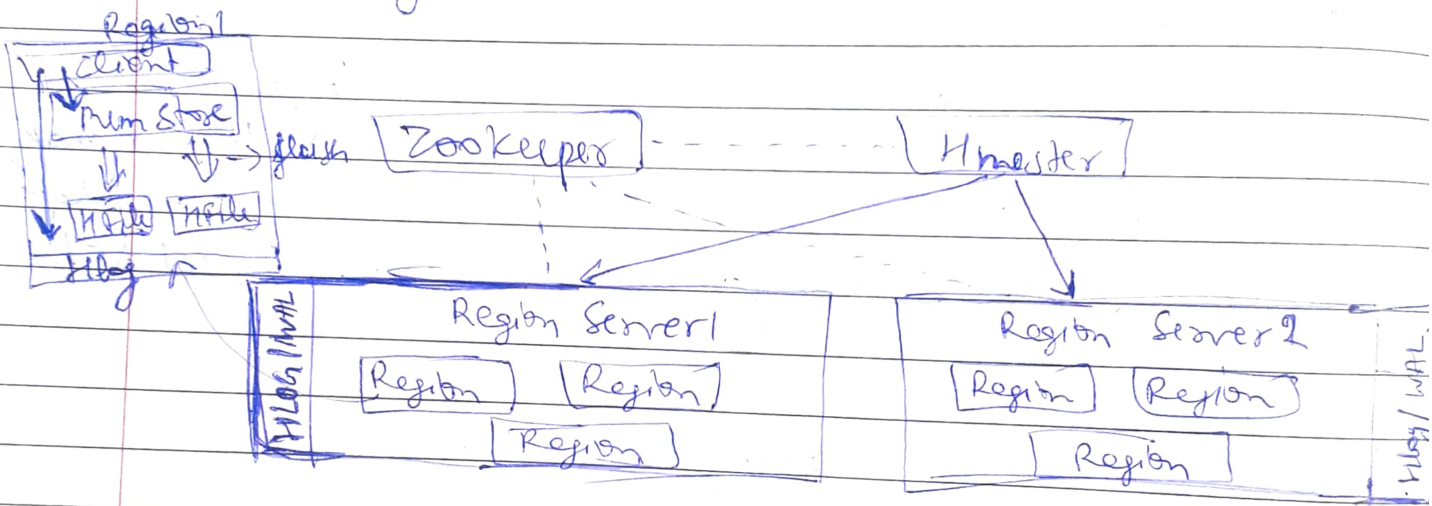
SQL → Rigid schema, consistency, transaction
NoSQL → speed, flexibility, scalability

Page No.
Date / /

Hadoop ke upar ka karne ke l banaya h

Hbase

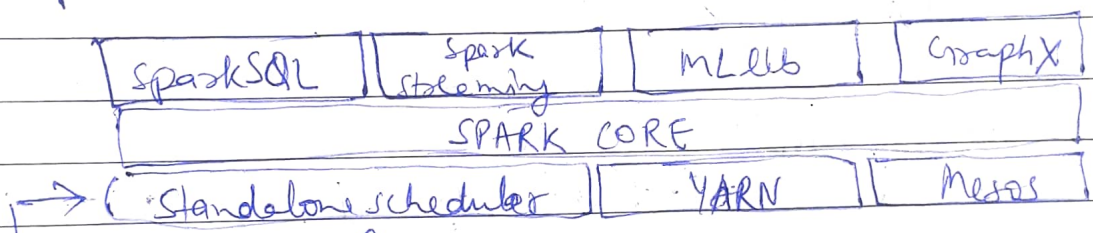
- NoSQL data store build on top of HDFS.
- column - Oriented data store.
- Store large amount of data (TB, PB)
- Big Data with random read and writes.
- handle various types of data.
- Horizontally scalable



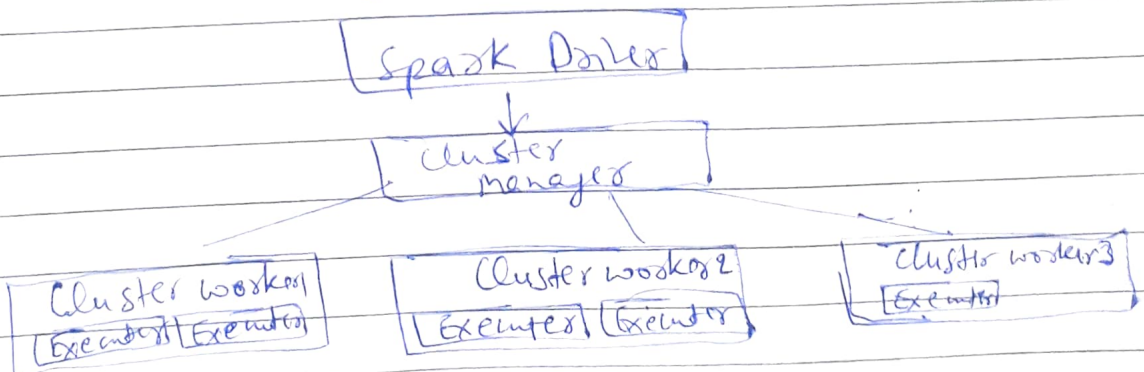
Hmaster - to monitor Region Server instances present in the cluster.
Regions Server - receives writes and read requests to a specific region, where the actual column family resides.
Region - consists of the distribution of tables and are comprises of column families
Zookeeper - centralized monitoring server which maintains configuration information & provide distributed synchronization

Apache Spark (Cluster computing platform)
- supports data analysis, ML, graph, streaming data etc
- can read/write from a range of data types and allows development in multiple languages.
- integrates closely with other Big Data tools.
- memory computation.
- Basic knowledge of Phyton, scala, SQL, Linux, Hadoop, statistics

Components



SPARK CORE - basic functionality of Spark (task scheduling, MM, fault recovery, interacting with storage systems)
SparkSQL - Spark's package for working with structured data.
Spark Streaming - Enable processing of live streams of data
MLlib - providing multiple types of ML algorithm.
GraphX - Library for manipulating Graphs.



SPARK ARCHITECTURE

Basics

- RDD - Resilient Distributed Dataset (: fault tolerant)
- DAG - Directed Acyclic Graph

RDD

- Fault tolerant
- Immutability
- Lazy Evaluation
- caching
- Partitioning

Transformations : ← Narrow T
← Wide T

```
var a = sc.textFile("hdfs://.../abc.txt")
var b = a.filter(...)        // RDD1
var c = b.distinct(...)      // RDD2
```

Actions
c.collect()

← Not take action till this point action on cluster

DOMS | Page No.
Date / /

Cassandra

It is a highly scalable, High-performance distributed database designed to handle large amount of data across many commodity servers, provide high availability with no single point of failure. It is a type of NoSQL database.

- Column-oriented database
- created at Facebook

Features

i) Elastic scalability - add more H/w to accomodate more data as per requirement.

ii) No single point failure

iii) Linear-scale performance - through put increases as you add no node, thus increase quick response time.

iv) Flexible data storage - accommodates all possible data formats include Structured, Semi-Structured and unstructured

v) Supports ACID property

vi)

DOMS | Page No.
Date / /

HIVE

- Data warehousing package built on top of hadoop.
- Used for Data Analysis.
- created for users comfortable with SQL.
- Query Language - HQL or HiveQL
- Used for managing and querying structured Data
- No Need to learn Java.
- Uses Map-Reduce for execution
- Developed at Facebook
- Not designed for OLTP and Does not offer real-time queries

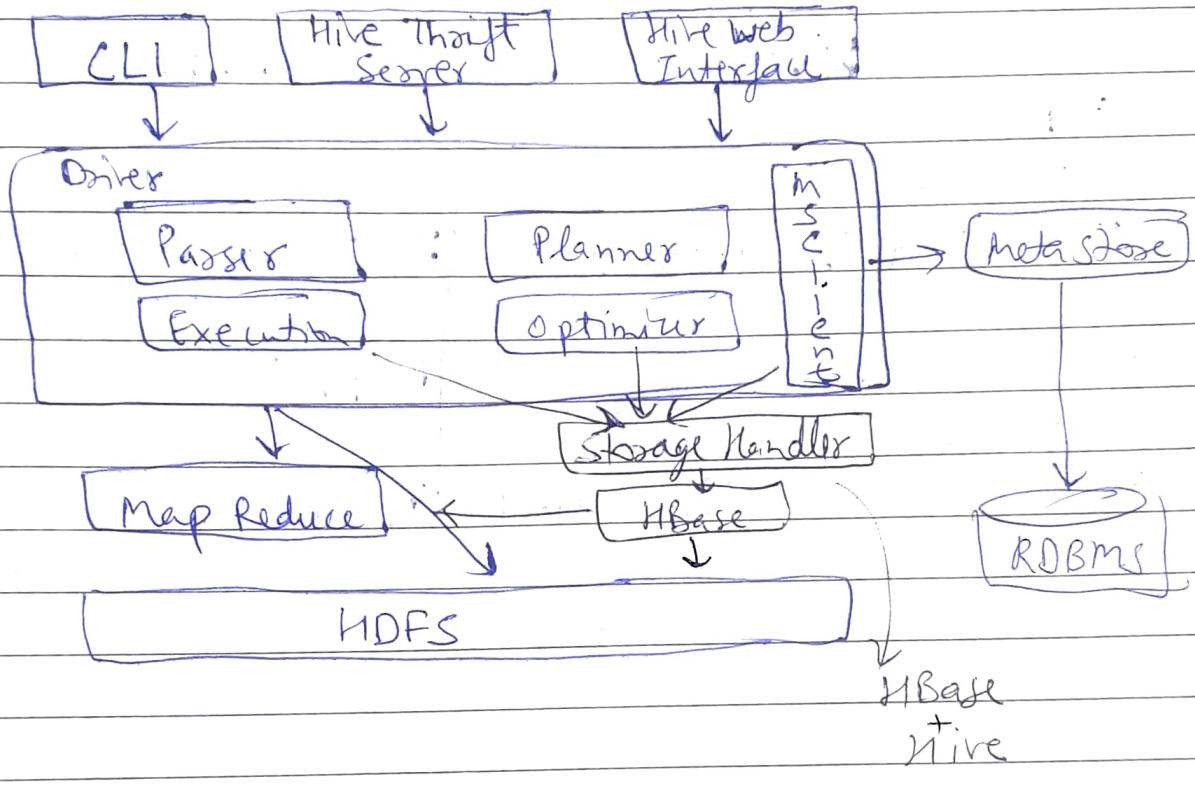
HIVE Features

- Creating Tables (Partitioning and Bucketing) ⇒ speed!
- JDBC/ODBC Drivers are available
- Data is stored on HDFS
- Uses hadoop for fault tolerance.

Where to use Hive

- Data Mining - to observes the pattern
- Document Indexing
- Predictive modelling - to predict the user behaviour
- Custom facing UI - to see in different way
- Spam Detection
- Ad Optimization

Hive Architecture



Integrating with HBase

i) Now single Hive query can now perform complex operation such as join, union.

ii) A lot of data sitting in HBase due to its usage in a real-time environment, but never used for analysis

Compare Hive & RDBMS

Hive	RDBMS
i) Focus on analytics and big data	i) Focus on online or analysis.
ii) Limited transaction support	ii) Transaction are usually support
iii) Distributed processing via map Reduce	iii) Distributed processing varies by vendor.
iv) Limited indexing support	iv) Full indexing support
v) No triggers	v) trigger support.

DOMS | Page No.
Date / /

Difference b/w HIVE, PIG, SQL

PIG

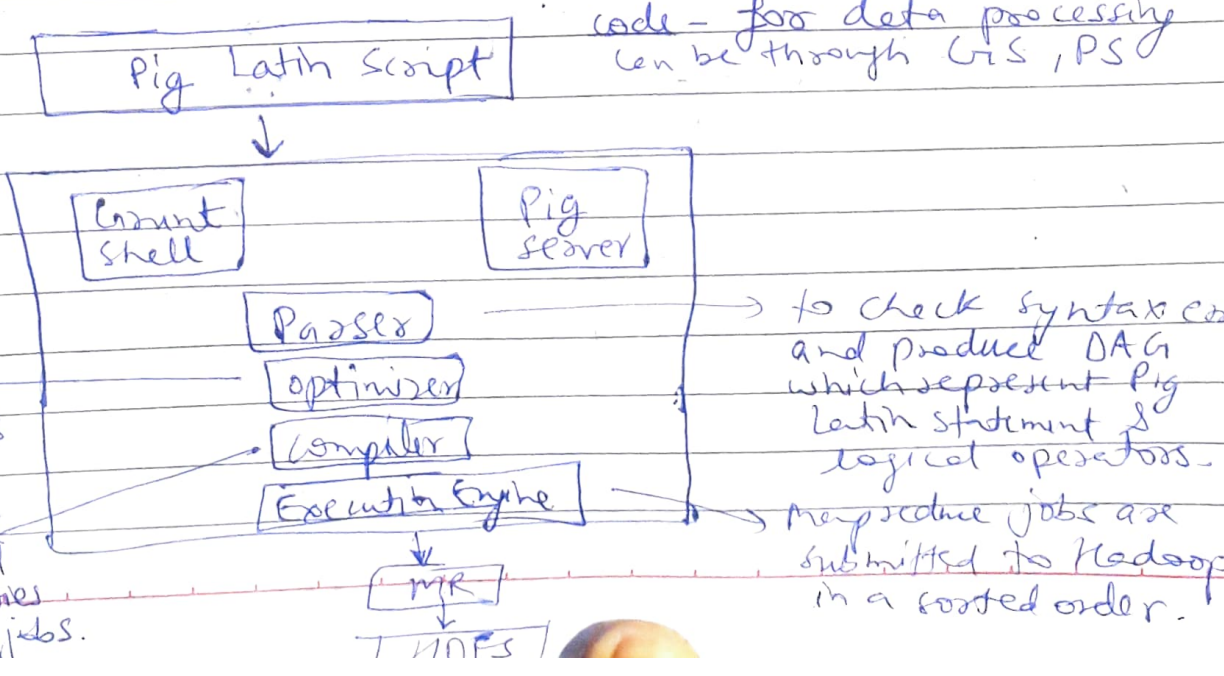
- Developed at Yahoo
- Apache pig raises the level of abstraction for processing large datasets.
- Pig is used to analyze larger sets of data by representing them as dataflows.
- works on top of the Hadoop

Why PIG?

- Programmer who are not so good with Java.
- Pig uses Multi-query approach.
- Built in operators :- join, filter, order
- Provides nested datatypes.

→ Pig is made up of two pieces
1. Pig Latin: - Used to express data flow.
2. Execution Environment - To run Pig Latin programs

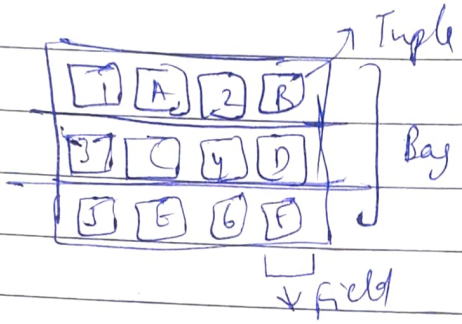
PIG Architechyse



Features of Pig

- Rich set of operators
- Easy to program
- UDFs
- Optimization
- Ability to handle all kind of data (eg - S

Data model



Apache Pig Vs MapReduce

Apache Pig	Map Reduce
• It is a Data Flow language	• It is a data proc paradigm
• HLL	• LLL and rigid
• Join operation is simple	• quite difficult
• basic knowledge	• Exposure with Jav is must to work with
• use multi-query approach ⇒ reducing the code size	• 20 times more li of lines of code.
• no need of compilation	• requires long compi

Pig	SQL
• procedural language	• declarative language
• scheme is optional	• schema is mandatory
• data model is nested relational	• data model is flat relational.
• limit -	• more opportunity for query optimisation.

	Pig	Hive
language used	• Pig Latin	• HiveQL
originally created by	• Yahoo	• Facebook
	• data flow language	• query processing language
	• procedural language	• declarative language
	• handle S, US, semi-S	• handle structured data

Application

- To perform data processing for search platforms
- To process time sensitive data loads.

Time Series

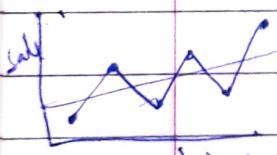
The set of data collected on the basis of time is called time series.

Importance

- Study Past Behaviour of data
- Forecast Future
- Estimates Trade Cycle
- Comparison
- Universal Utility

Components of Time Series

a) Trend - is a long-term increase or decrease in the series of time which persists over a long time.
It can be linear or Non-linear



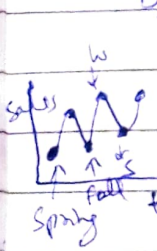
Deterministic trend

Trend in which the value of the time-dependent variables increases or decreases inconsistently.

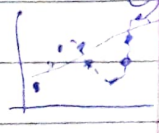
Non-deterministic trend

Trends in which the value of the time-dependent variable increases or decreases inconsistently.

b) Seasonality -
It is the regular pattern of up and down fluctuations in a time series. It may be a short term variation occurring due to seasonal factors.



c) cyclicity - It can be defined as a variation caused by circumstances that repeat in irregular intervals.



d) Irregularity - variation occurs due to unpredictable factors and also do not repeat in particular patterns.

Time series Model

Non-seasonal Model with Trend

$X_t = m_t + Y_t$

- X_t: Stochastic process
- m_t: trend
- Y_t: random noise

Basic methods to estimating/eliminating trend –

1. Trend estimation
   - by Regression Analysis
   - by Linear Trend Forecasting
2. Trend elimination by differencing

Methods

1. Filtering
2. Differencing
3. Joint fit method

Classical Decomposition Model with Trend

$X_t = m_t + S_t + Y_t$

- S_t: seasonal component

Method of Measuring Trends

1. Freehand Curve Method
2. Semi-Average Method
3. Moving Average Method
4. Least Squares method.

DOMS | Page No.
Date / /

Data Science

Theories and techniques from many fields and disciplines are used to investigate and analyze a large amount of data to help decision makers in many industries.

Data Science	Data Analytics
• It is a field that refers to the collective processes, theories, concepts, tools and technologies that enable the review, analysis and extraction of valuable knowledge and information from raw data.	• Data Analytics is the process of examining data sets in order to draw conclusions about the information they contain, increasingly with the aid of specialized systems & S/W.
Tools & Languages • Python, SAS, SQL	• R, Tableau Public, Apache Spark
Application Areas • Digital advertisements, Recommender systems	• Health care, Travel

Scientific Modeling	Data-Driven Approach
• Physic-based model	• General intrinsic engine replaces model.
• Problem-structured	• Structure not related to problem
• Mostly deterministic, precise	• Statistical models handle stochastic randomness, and un-modeled complexity

Ravinder
COE-1

Data Mining (refer AB)

DOMS | Page No.
Date / /

Machine Learning	Data Science
• Develop new (individual) models	• Explore many models, build and tune hybrids.
• Prove mathematical properties of models	• Understand Empirical properties of model
• Publish a paper	• Take action

	Data Science	Data Engineering
Approach	• Scientific (Exploration)	• Engineering (Development)
Problems	• Unbounded	• Bounded
Path to solution	• Iterative, exploratory, non-linear	• Mostly linear
Presentation skills	• Important	• Non Important
Research Experience	• Important	• Non as important
Programming skills	• Not as important	• Important
Data Skill	• Important	• Important