What is a data stream?

→ A stream is defined as a possibly unbounded sequence of data items or records. That may or may not be related to, or correlated with each other.

e.g instruments, many IOT appln areas, computer programs, websites or social media posts.

→ ~~datas~~ each datas are timestamped and in some cases geo-tagged.

- streaming data sometimes → referred to as event data → each data item is treated as an indv. event in a sync sequence.

Synchronized sequence of events:

~~Data stream~~

Streaming Data Systems

Quiz
- Manage one record or small time window
- Near-real-time
- Independent computations
- Non-interactive

Dynamic steering ← part of streaming data management & processing.
→ self driving cab.

= ~~data streamin~~

- Some streaming Data ~~Stream~~ systems
- amazon kinesis.
- Apache storm
- Flink
- spark Streaming
- Samza

Why is steaming Data different?

Data-at-rest.
- mostly static data from one or more sources.
- collected prior to analysis.

Data-in-motion
- analyzed as it is generated.
- stream processing. ← analysis of stream data.

~~Data Processing~~ ← ~~analysis of~~

Data Processing Algorithms:

Static/Batch $\xrightarrow{\text{Processing}}$ size determines, time and space

Streaming Processing → Unbounded size, but ~~infinite~~ finite time & space.
Quiz

Streaming Data Management & Processing:

- compute one data element or a small window of data elements at a time.

ACID & BASE

↳ difficult to maintain in a DBMS.

↳ relaxes ACID

↳ BA: Basic Availability

↳ S: Soft state

↳ E: Eventual consistency

CAP Theory: A distributed computer system cannot simultaneously achieve.

- consistency
- Availability
- Partition Tolerance

The

Week 4 Imp

Schema on Write	Schema on read
(i) data is structured ~~stored in~~ & enforced to adhere to a specific schema before being written to database / datawarehouse	(i) data is stored without any predefined structure or schema.
(ii) ensures data integrity & consistency	(ii) schema is applied at the time of reading
	(iii) allows more flexibility in handling diverse & unstructured data.
(iv) used in datawarehouses	(iv) used in data lakes

write vice versa

AsteriskDB ← apache

semis
↳ used to deal with ~~[illegible]~~ structured data.

↳ fully-fledged, provide ACID properties like data integrity

↳ structured to deal with data that doesn't fit in rows & columns. like JSON

↳

act as namespace for data

working

↳ organizes data into structures called dataverse & data types

define structure of data

↳

e.g. tweets → nested parts like entities & user information } AsterixDB explores hierarchical structure in its schema.

↳ handle geospatial data too, geographical info

↳ uses a query language. ~~like~~ ~~XML~~ similar to XLL called AQL.

↳ enables querying multiple languages like XQuery, Hadoop, Mapreduce, etc.

↳ operates on clusters of machines.

↳ divides data into partitions & executes queries

→ These computations can update metrics, monitor, & plots statistics

- Relatively fast & simple computations
- No interaction with data sources.

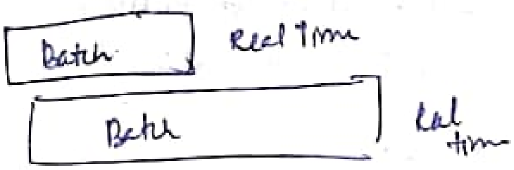
Hybrid architecture

↳ lamda architecture - for processing streaming & batch Jobs at the same time.

→ In these systems streaming data elements are pushed to a batch system and become available to access & process as batch data.

In such systems, stream storage layer is used to enable fast trees of streams & ensure data ordering & consistency.

λ architecture



streaming data changes over time.

size + frequency.

- size → unbounded
- size & freq → Unpredictable
- Processing → Fast & simple

Changes can be periodic or sporadic

Periodic : evenings, weekends, etc.
Sporadic : major events
} Quiz

other changes include dropping or missing data or even no data

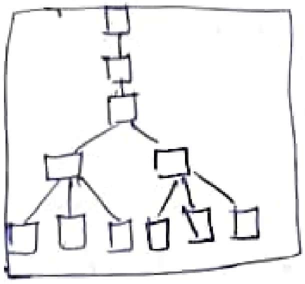
Data Lake ← big data storage & processing challenges.

→ part of data structure that many streams can flow into & get stored for processing in their original form

Quiz

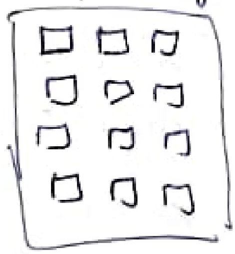
Data Warehouse vs Data Lake

↓ Hierarchical file system



↓ Structure format
↳ when data is in use, then stored in warehouse

↓ object storage



↓ raw format
↳ data is stored as it gets streamed.

→ Schema-on-write → Schema-on-read.

Data Lake Object Storage

(i) Each ~~object is~~ data is stored in binary large object (BLOB) & is assigned a unique identifier.

(ii) Each data object is tagged with a # of of metadata tags.

↑ data is search using the tags.

How a Data Lake Works

→ Load data from source
→ Store raw data
→ Add data model on read

classmate
Date
Page

Programming Models on Big Data

→ an abstraction or existing machinery or infrastructure.
→ set of run time libraries + programming libraries
→

requirements: (i) Support big data operations → split ~~data values~~ volumes of data
→ Access data fast
dis tr. computation to nodes

(ii) Handle fault Tolerance → Replicate data partition
→ recover files when needed.

(iii) Enable adding more racks
(iv) Optimize for specific datatypes

Hadoop Goals

(i) Enable scalability
(ii) Handle fault tolerance
(iii) Ability to handle diff types of data
(iv) facilitate a share env.
(v) Provides value → community supported
→ wide range of Application

Hadoop Ecosystem

(i) HDFS → scalability to large data sets
→ reliability to cope hardware failure

2 components
(a) Name nodes → keep track of filename, location in directory, etc.
→ mapping of content in datanode
(b) Data nodes → stores file blocks
→ Listens to namenodes for block evaluation, deletion & replication.
→ replication is done for fault tolerance & data locality.

YARN: flexible scheduling & resource management over HDFS.
Hive & Pig - ~~A simplifies parallel computing~~
- ~~Your only need to given MapReduce two function~~
~~Map → applies~~
~~Reduce → summarize ()~~

Note: Data computation framework → resource manager & node manager

classmate
Date
Page

Cloud Services → Service Models [Application / Platform / Infrastructure]

IaaS = Get the Hardware only, Amazon EC2 cloud.
↳ bare min^{m} rental service

PaaS = - Platform as a service, google App engine
- Get the computing env. Microsoft Azure

SaaS = Get full s/w on demand
- service model
- dropbox.

Decision depends on
- skill
- demand
- capital
- Security

XaaS = Anything as a services.
- Storage aaS
- Marketing aaS
- Communication aaS.

Hadoop layer diagram:
Low level i/f → ~~data~~ storage & scheduling, on the bottom.
High level language & interactivity at the top.

Yahoo
SQL-like queries — data flow scripting
facebook
Hive & Pig: augment data modeling of MapReduce with relational algebra & data flow modeling respectively.
Giraph: ~~out~~ large scale graph efficiently

Storm, Spark, Flink: used for real time & in-memory processing of big data.

Hbase, cassandra, MongoDB: No SQL for non-files

Zookeeper: management like synchronization, configuration & high-availability.

indexing
Map Reduce: relies on YARN to schedule & execute parallel processing over distributed file system intents.
Map → Apply operation
Reduce → Synchronize operation.
Map → Shuffle & sort → Reduce

* Where to use hadoop
→ Many platform over single data sea
→ High volume
→ High variety
→ Random data Access, Infrastructure replacement

Where not to
→ small dataset
→ Advanced algo
→ Task parallelism

Cloud Computing & cloud service
(i) Build Resource
(ii) Cloude

classmate

Date
Page

Steps in data science Process

1. Acquire data
   - ~~Identify~~ determine data set ^ to look when to look ~~data~~
   - ~~Retrieve data~~ Suitable data
   - Query data ⟶ Relational data, R,
2. Prepare data
   - → Explore (Preliminary analysis, Understand nature of data)
   - → Pre-process data - (Clean, Integrate, Pkg)
3. Analyze Data
   - → Select analytical techniques
   - → Build models.
4. Communicate Results
5. Apply Results.

Acquiring data.
- identify suitable data
- acquire all available data.

Webservices { REST → Representational State Transfer
SOAP, WebSocket

NOSQL storage - | mongodB, cassandra, neo4j, Apache, couch DB |

↕ API ← provide API allows user to access data using

↕ Webservice ← using REST

Webservice

eg → Acquiring Data from WIIDFIRE

Historical Weather ⟵ SQL
Curr. Weather ⟵ WebSocket
Real-time tweets near fires ⟵ REST

Tradⁿ dB → SQL & query browser
Remote data → Webservices
Text files → Scripting Languages
No SQL ~~dat~~ storage → Web Services, Programming Interfaces

classmate
Date
Page

5 P's
Components of Data Science

People } → Process } Programmability | metrics → Product.
Purpose }   ↓ Platform }

Big Data Engg.
Acquire > Prepare >
└ scale ↑ └ scale ↑

Computational Big Data science
Analyze > Report > Act >
└ scale ↑ └ scale ↑

Process: Build metrics for accountability
Cost
Timeline
Planning of deliverables
Expectations
Purpose

Asking the right question:

Assess situation → Risks, Benefits, Contingencies, regulations, resources, requirements
Define Goals → Objectives
→ Criteria

Steps to find right problem to tackle in data science
Define Problem
↓
Assess situation
↓
Define Goals

} formulate the Question

→ Explore

Goal: To Understand Data

explore dependencies of ~~variable~~ b/w diff variables of data

Why explore?

sales prices going up or down.

to detects errors

Correlation   General trends   Outliers

data pt distants from other data pt.

Describe your data

Mean & Median ← represent the ~~distance~~ location of a set of values.

Mode ← value that occurs most frequently

range & standard deviation → measure of spread in your data

Visualize your data
- Histograms
- Line Graphs
- Heatmaps
- Scatter plots
- Box plots

Step 2B - Prep-processing data

clean + transform

data quality issues
- inconsistent values
- Duplicate records
- Missing values
- Invalid data
- Outliers

Addressing Data Quality Issues.

Remove data with missing values

Merge duplicate records

Generate best estimate for invalid value

Remove Outliers

Domain Knowledge

classmate
Date
Page

Valence: Measure of Connectivity.

- Data connectivity → two data items are conn. when they are related to each other
- Valence - fraction of data items that are connected out of total # of possible connections.

Note: Valence increases over time
Makes the data connections denser

Challenges:
- More complex data exploration algo.
- Modeling & prediction of valence changes.
- Group event detection.
- Emergent behaviour analysis.

?

$\boxed{\text{Strategy} \in \text{Big data}}$
- Aim
- Policy
- Plan
- Action

1st step in determining a big data strategy

→ Business Objective → ← Long term
← Short term

Provide organizational buy-in
- commitment, • sponsorship, • communication.

Build diverse teams: • Diverse expertise, • deliver as a team.

Share data:
- Remove barriers to data access
- No data silos ← big heap or tall tower.
- Data sharing mindset

Define big data policies: Privacy & Lifetime (whom should be given access)
Curation & quality. data curation & quality
Interoperability & regulation.

are these ... of ... data

Cultivate analytics-driven culture:

Taking all the decisions using analysis of data

Analysis + Business = Opportunities

Integrate analytics → Comm. unicate goal → build team → Share data → Adopt for new situation → Integrate analytics

classmate

Date

Page

Velocity: Speed $\frac{\Delta x}{\Delta t}$

Speed of creating, storing & ~~str~~ analyzing data.

- Real time processing ← gathering weather info for travel
  ← sensors saving lives.

Batch Processing.

Collect Data → Clean Data → Feed in chunks → Wait → Acts } incomplete

Real-Time Processing

Instantly capture data → Streaming → Feed real time to machines → Process Real Time → Act } fast

Rate needed for data-driven actions

↓

Rate of genⁿ & processing of data.

Streaming data ⇒ "what's going on right now" + streaming data ⇒ gets generated at a varied rates

$\boxed{\text{Real-time processing}}$

Agile & adaptable business decisions

Veracity: Quality - validity, volatility

- Accuracy of data
- Reliability of the data source
- Context within analyses.

Week-2 & 3.

Week-2

4 v's

Characteristics of Big Data
- ~~Volo~~ Volume — vast amt of data that is generated every second, etc.
- Variety - diff forms that data can come in
- Velocity - pace at which data moves from one pt to the next
- Veracity - refers to biases, noise & abnormality in data
- Valence - connectedness of data

Volume:

Challenges:

Storage $\xrightarrow{\text{distr}}$ Processing

performance

volume ↑ ↓ ↑ cost

Volume = size → Challenges
- storage
- Access
- Processing

Variety:

Axes of Data Variety

diff in the representation of the data → Structural Variety - formats & models — satellite images in wildfire very diff from tweets sent by ppl.

EKG signal to newspaper article

Semantic Variety — how to interpret & operate on data.

we use qualitative vs quantitative measures — age can be a # or we can rep as juvenile, infant

Media Variety - medium in which data get delivered. ← medium in which data gets delivered

eg - audio of a speech & transcripts.

Availability variations -
- real-time ? ← traffic among
- Intermittent ? ← satellite data

Scalability issue   Impact of data variety:
- harder to ingest, • difficult to create common storage.
- difficult compare & match data across variety, • diff. to integrate
- management & policy challenges.

Week-3 File System → long term info storage system
→ It can access result of process data
→ store large amt of info
→ enable access of multiple process.

Date
Page

Distributed File System: (DFS):
→ replicated the data b/w the racks & also computer across the geographical location.
→ DFS makes the system more built tolerant,

High concurrency vs Low Consistency

Data Partitioning
Data replication
↓
Data scalability
Fault tolerance
High concurrency

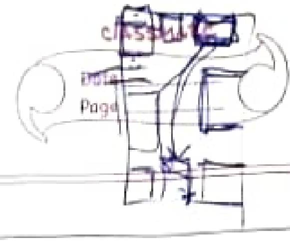
Scalable Computing over I/n.

Commodity Clusters:- are affordable parallel computer with avg # of computing nodes conn. to each other via fast n/w.
→ less specialized.
reduce computing cost.

Distributed Computing → computing in one or more of these cluster across a local area n/w on Internet.
enables data parallelism.

Common failure in commodity cluster:
- failure of entire rack.
- failure of connection b/w rack & n/w.
- failure of connection b/w two ~~each~~ nodes in ~~the~~ a rack.

Failure ⇏ Complete Restart
↓ ↓
Redundant data storage
Data parallel Job restart



Getting Data in Shape

also called Data Munging, Data Wrangling, Data Preprocessing.

Data Munging:

- when dataset has large # of dimension → Dimensionality Reduction (3D to 2D)
- raw data, manipulated to be in correct format for analysis ← Data Manipulation
- Transformation ← reduce noise & variability.
- comparing height & weight, weight in much you 150, scaling will equalize contribution — Scaling — scale values b/w zero & one
- Feature Selection — remove, combine, add features
  - remove: redundant & irrelevant features
  - add: feature

Categories of Analysis Techniques:

predicted vs correct values:

(i) Classification ← Predict category

(ii) Regression ← Predict numeric values ← sales, stock price

(iii) Clustering ← Organize similar items into groups.

investigate & validate:

(iv) Association Analysis ← Find rules to capture associations b/w items

(v) Graph Analytics: Use graph structures to find connections b/w entities

Modeling

Select Techniques → Build Model → Validate Models

Communication

present using visualization tools.

R, Python, D3, Leaflet,

Tableau, Google Charts, Timeline

- Tableau ← create visualizations in your public profile
- Google Charts ← cross platform compatibility
- Timeline ← allows to create timeline.

Action:

Process ⟷ Automation ⟷ stakeholder

Assess Impact:

Monitor

Measure

Evaluation