

# Intro to Data Engineering

What do Data Engineers do?

6 V's of Big Data

Where do we actually store data? How do we separate data storage based on use cases?

What is an ETL pipeline?

Understanding Instance, Node, Cluster

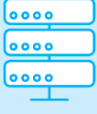





Multiple types of Distributed clusters?

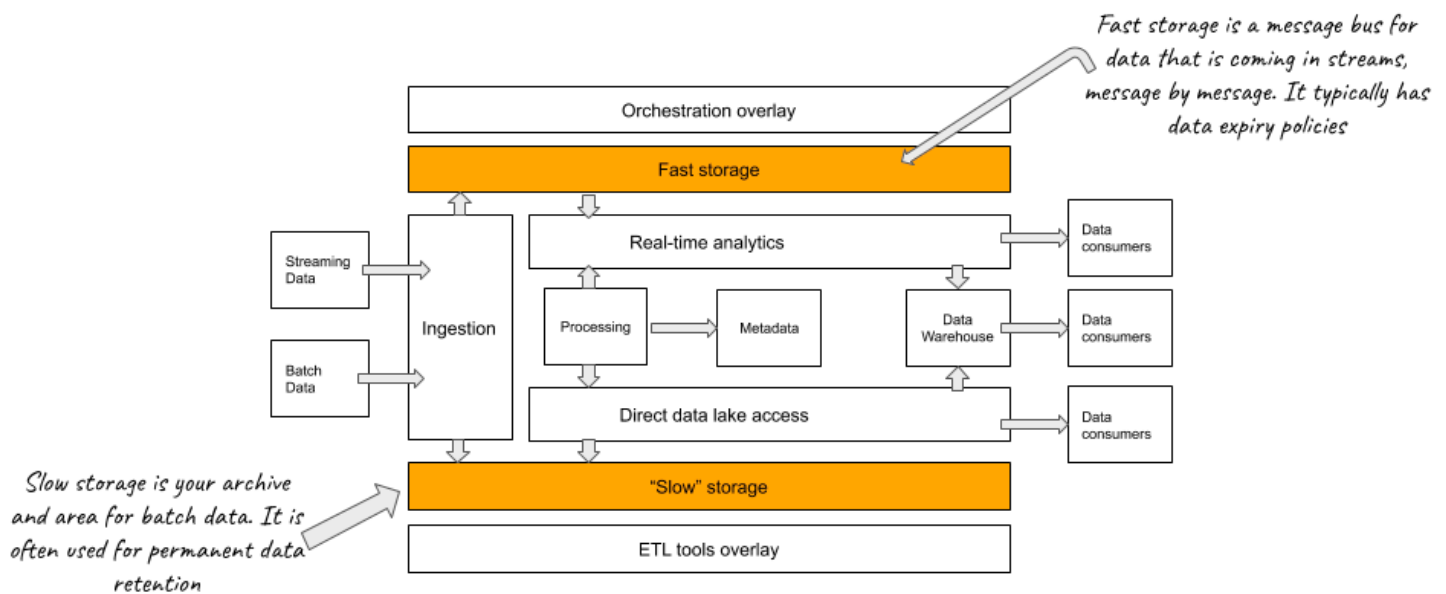
---

HOW'S THE  
JOSH?

# The six Vs of Big Data

Big data is a collection of data from various sources, often characterized by what's become known as the 3Vs: volume and velocity. Over time, other Vs have been added to descriptions of big data:

VOLUME	VARIETY	VELOCITY	VERACITY	VALUE	VARIABILITY
The amount of data from myriad sources	The types of data: structured, semi-structured, unstructured	The speed at which big data is generated	The degree to which big data can be trusted	The business value of the data collected	The ways in which the big data can be used and formatted
					



1. **SQL/NoSQL Database** : Playing with SQL/NoSQL DB and evaluating their use cases.
2. **Data Modeling and Data Warehousing** : Focus on creating fact,dimension tables, SCS with cons and pros of storage types
3. **Batch pipeline** : Showcasing how big giants perform their reporting, analyze data on dashboards and then making data driven decision
4. **Real time Streaming Pipeline**: Creating our own crypto dashboard tracking bitcoin fluctuations.

5. **Orchestration with Airflow:** once we have relevant workflow, coordinating the execution and monitoring of these workflows will be done via airflow
6. **Git and github :** Pushing our code to the version control system to keep track of code changes and to collaborate on code.
7. Intro to Cloud, covering all major ones
8. Databricks/snowflake hands on

## Rules :-

① class is going to begin at sharp

9:02 PM.

② class content will be covered till  
11:00 PM (11:00 to 11:30 PM)  
DCS

③ Use chat box for immediate doubts,  
use Questions tab for asking  
anything who is related to  
course or even not related  
to course.

\* ④ Solve the Assignments before coming  
to next class.

15 Days / month  
↳ Lab access

⑤ Any feedback related to class  
is always welcome. (5)

⑥ All major concepts are going  
to be revised by default 2 times,  
slightly

3 times revise of it is 0-0-0  
complex.

## Data Engineering

↳ is the process of designing and building systems to collect, store and analyse large amount of data.



Data Engineers	Data Scientists
<p>① DE builds the pipelines that collect and deliver data for DS.</p> <p>② Skills :-</p> <ul style="list-style-type: none"><li>① Infrastructure Components → (VM, NW, LB) ... (IaaS/PaaS)</li></ul>	<p>① DS are the people who analyze data, create algorithms and make Prediction on that data.</p> <p>② Skills</p> <ul style="list-style-type: none"><li>① Programming Languages:- R, Python, SQL</li></ul>

- (2) Cloud (AWS/Azure)
- (3) DB and DWH  
(MySQL/Oracle...  
DB/Redshift)
- (4) Proficiency working with  
Data Pipelines  
 (a) Beam  
 (b) Airflow
- (5) ETL Tools  
↳ Informatica
- (6) Languages of  
→ SQL  
→ Python or Java  
or Scala
- (7) Big Data processing  
tools like  
databases, Spark, Hadoop  
Snowflake

- (2) Strong mathematical  
Knowledge:-  
 (a) Statistics & Prob  
 (b) Linear Algebra &  
Calculus  
 (c) Matrix
- (3) ML & AI Libraries  
 (1) Scikit-learn  
 (2) Tensorflow  
 (3) Pytorch
- (4) Big Data technologies:  
 (1) Spark  
 (2) Hadoop
- (5) Databases:-  
 SQL/NoSQL  
 ↓  
 Oracle  
 MySQL  
 MongoDB  
 Cassandra  
 Redis
- (6) Cloud

Challenges (6Vs)

- ↳ Volume
- ↳ Velocity
- ↳ Variety
- ↳ Veracity
- ↳ Value
- ↳ Variability

① Volume: Amount of data that a DE needs to store.

② Variety:

✓ Structured	✓ Semi-Structured	✓ Unstructured																																
<p>↳</p> <table><tr><td>c1</td><td>c2</td><td>c3</td><td>c4</td></tr><tr><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td></tr></table> <p>⇒ DBS</p> <ul style="list-style-type: none"><li>↳ MySQL</li><li>↳ Oracle</li><li>↳ MSSQL</li><li>↳ Postgres</li></ul> <p>→ (eid int)</p>	c1	c2	c3	c4													<p>→ loosely organized into categories using meta tags.</p> <p>→ CSV, TSV, JSON, XML, Parquet, URC</p> <p>↓ Avro, Yaml</p> <table><tr><td>c1</td><td>c2</td><td>c3</td><td>c4</td></tr><tr><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td></tr><tr><td></td><td></td><td></td><td></td></tr></table>	c1	c2	c3	c4													<p>→ Data which never follows a defined format</p> <p>→ Videos</p> <ul style="list-style-type: none"><li>↳ Images</li><li>↳ Audio</li><li>↳ PDF</li><li>↳ Text...</li></ul>
c1	c2	c3	c4																															
c1	c2	c3	c4																															



③ Velocity :- Speed at which data is getting generated.

Batch	Streaming (Runtime)
→ DE already knows the size of data and then process it.	→ 24x7 allows a DE to feed the data into analysis tools and get instant insight.

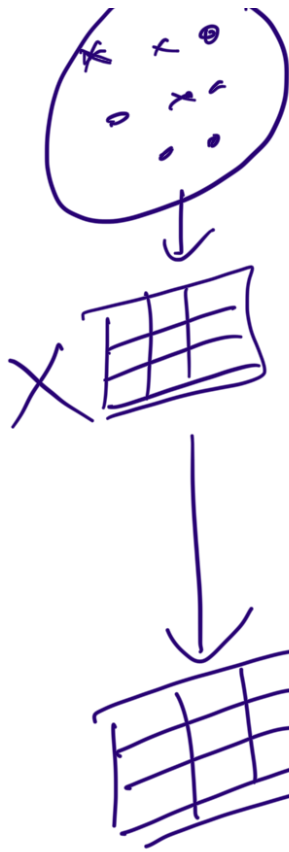
④ Veracity :- the accuracy, reliability and trustworthiness of data.

⑤ Value :- the usefulness of data in decision making.

⑥ Variability :- the changing nature & inconsistencies in data.

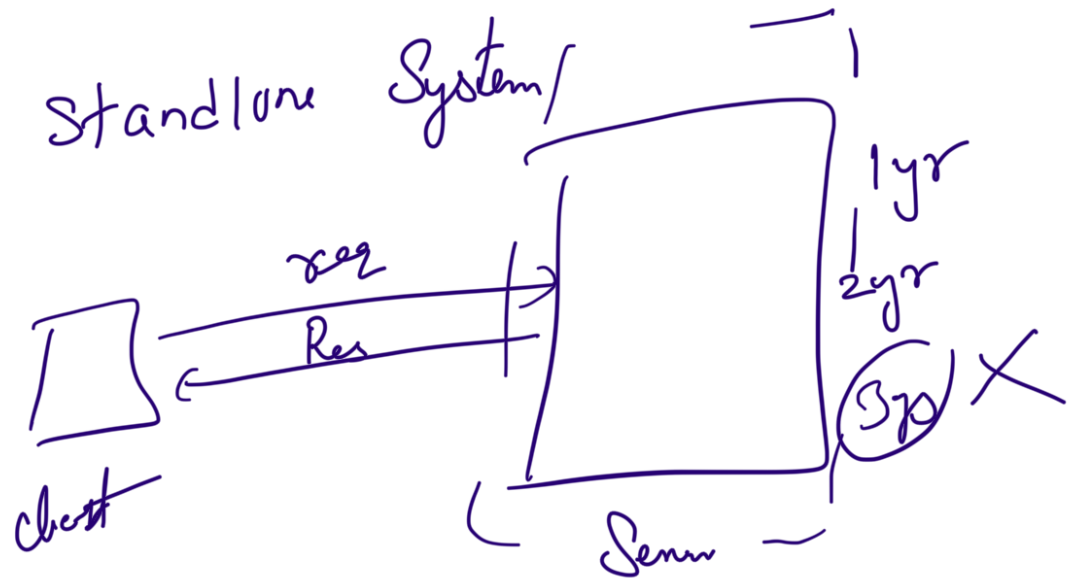
ETL

4 Extract → Data is pulled from that are



4 Transform  
 ↳ data gets changed, mapped & transformed to meet operational needs.  
 to get sources  
 usually heterogeneous such as OLTP, data warehouse

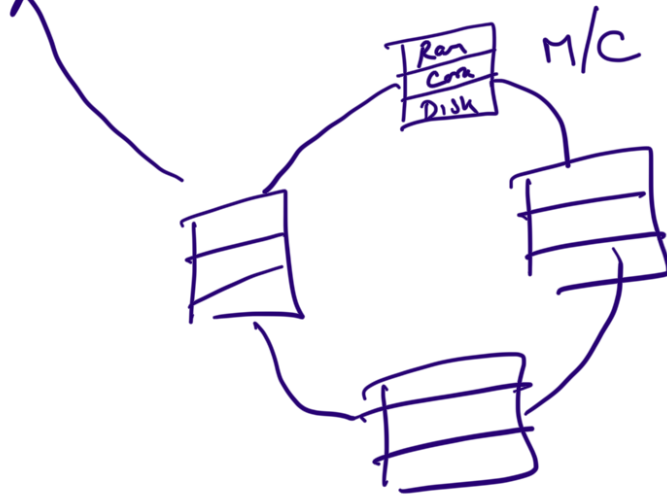
4 Load  
 ↳ Process of writing converted data from a staging area to a target system.



Distributed System :-

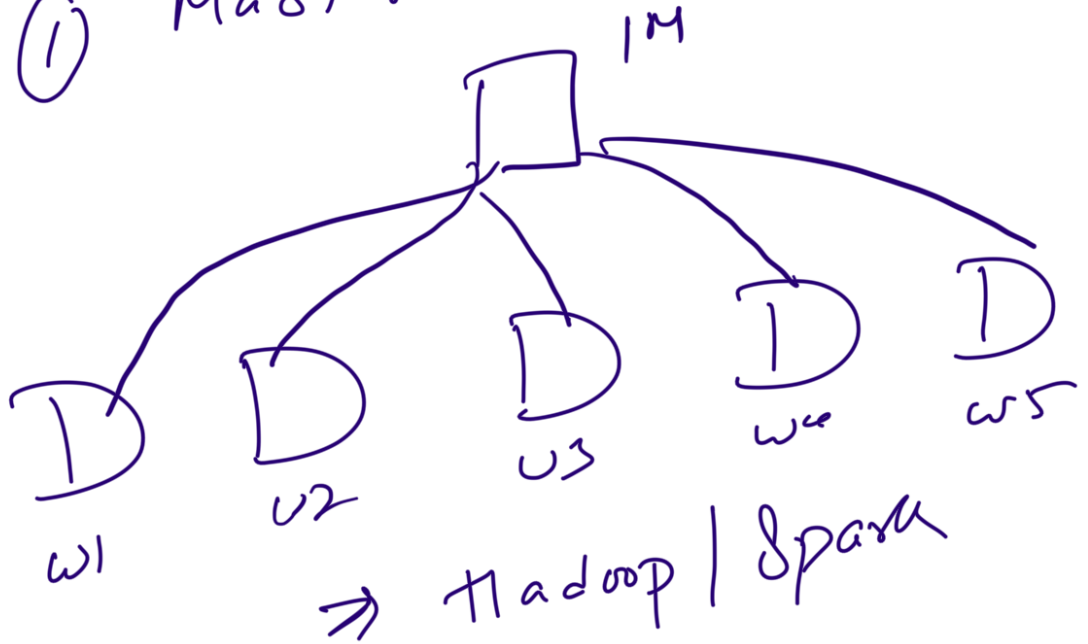
Hb = 40,000 ✓  
 Apple = 60,000 ✓  
 NF = 50,000 - 80,000

✓  
Cluster / node / Instance → H/W machine  
VM cloud



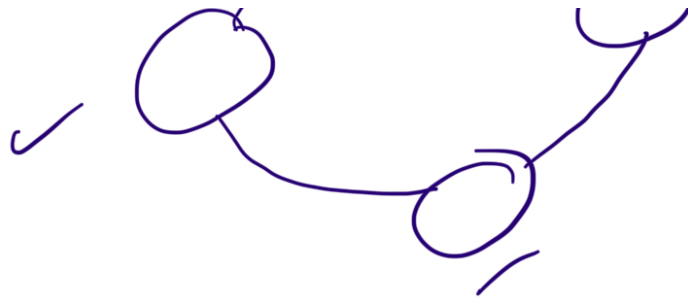
Distributed cluster

① Master worker



② Peer to peer X





C\*

OLAP	OLTP
<ul style="list-style-type: none"> <li>→ online Analytical Processing</li> <li>→ Stores data for longer period of time,  <div style="border: 1px solid black; border-radius: 50%; padding: 5px; display: inline-block;">           5y, 10y, 20y or         </div> </li> <li>→ Usually used to represent Dashboard, Reports, Analysis</li> <li>→ Read intensive sources</li> <li>→ Ex: DWH → SSIS  <div style="margin-left: 40px;">↓</div> <div style="margin-left: 40px;">Redash</div> <div style="margin-left: 40px;">Bijgun</div> </li> </ul>	<ul style="list-style-type: none"> <li>→ Online txn Processing</li> <li>→ Stores data for shorter period (1yr)</li> <li>→ usually used to store day to day trans.</li> <li>→ write intensive sources</li> <li>→ Databases: mysql  <div style="margin-left: 40px;">↳ Oracle</div> <div style="margin-left: 40px;">↳ Postgres</div> <div style="margin-left: 40px;">⋮</div> </li> </ul>

