DSBDA - U6

Histogram :

→ A histogram is a type of bar chart that represents distribution of numerical data by grouping values into ranges called bins.

→ Width of bar is decided by the numeric proportions of grouped data.
Height is decided by the freq.

→ Eg.: Consider the data of test scores:
14, 10, 12, 19, 18, 20, 10, 8, 5, 0, 2, 9

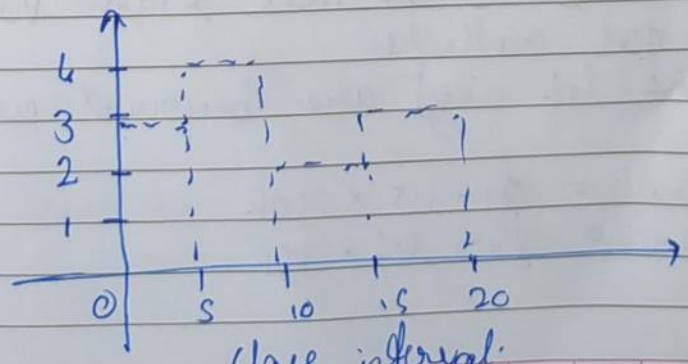
① Step (1) : In ascending order

[0, 2, 5, 8, 9, 10, 10, 12, 14, 18, 19, 20]

② Step (2) : Div. into bins, decide the range.

Class Interval	0-5	5-10	10-15	15-20
Freq.	3	4	2	3

③ ~~Plot~~ Step (3) : Plot



↳ Usage:

① Data distribution: Understand how data pts. spread over an interval.
② Detect skewness: Identify whether data is skewed or symmetric
③ Outliers: Bins w/ very low/high freq. or isolated may indicate outliers.

Box plot:

↳ A box plot is a graphical repsn of data that shows central tendency, spread and potential outliers

↳ Summary stats. viz. in box plot are:
- Minimum
- First quartile (Q1) 25%
- Median / Second quartile (Q2) 50%
- Third quartile (Q3) 75%
- Maximum

↳ Characteristics:

- Box shows Interquartile Range (IQR) = Q3-Q1
- Vertical line in box shows median (Q2)
- Whiskers extend to max & min. values that are not outliers.
- Outliers (if any) are beyond whiskers.

↳ lower limit = Q1 - 1.5 × IQR
upper limit = Q3 + 1.5 × IQR

↳ Eg.:

Consider data = 100, 120, 110, 150, 110, 140, 130, 170, 120, 220, 140, 110

① Ascending order:

[100, 110, 110, 110, 120, 120, 130, 140, 140, 150, 170, 220]

② Q1, Q2, Q3:

Q2/median = $\left(\frac{n}{2}\right)^{th}$ (odd) · OR $\left(\frac{n^{th} + (n+1)^{th}}{2}\right)$ (even)

Median = $\frac{120+130}{2} = 125$

Q1 = First 6 vals ka median. If odd, exclude Q2/Median

Q3 = Last 6 vals —, —

$Q1 = \frac{110+110}{2} = 110$ | $Q3 = \frac{140+150}{2} = 145$

IQR = Q3 − Q1 = 145 − 110 = 35

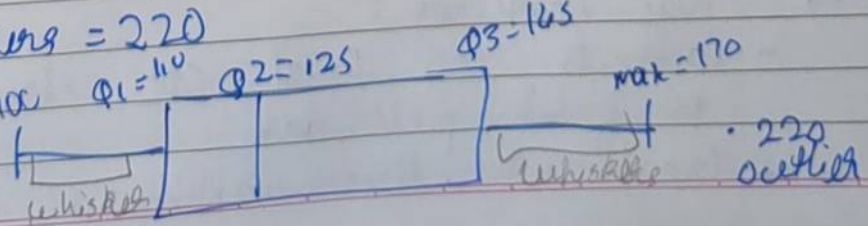
lower limit = Q1 − 1.5 × IQR = 57.5

upper limit = Q3 + 1.5 × IQR = 197.5

∴ min, max bet^n [57.5, 197.5] are 100, 170 resp.

Outliers = 220

Repr^n :



↳ Usage:
- Identify outliers
- Summarise large datasets
- Compare distributions of diff. grps.
- Skewness of data (if median not centered)
- Whiskers help identify spread of data.

Diff. of histogram from density plot:

↳ A histogram repr^n ~~does~~ freq. of bins whereas density plot repr^n probability density funct^n.

↳ Uses bars (hist)	Uses continuous curve (density)
↳ No smoothing; depends on bin width	Smoothed repr^n
↳ Area of bars repr^n count	Area under the curve = 1 (probab)
↳ Outliers less visible	Outliers easily identified.

MS PowerBi:

↳ - It is a powerful BI & data viz. tool by MS.
- Enables users to connect multiple data sources, transform raw data, create interactive dashboards & reports
- Easy to use & simple GUI
- Uses based sql^n
- Fast & efficient

↳ Key features:
- Connects w/ Excel, SQL, cloud services, APIs to fetch/pull data.
- Create interactive dashboards & reports

- AI integration for AI-driven insights and natural lang. queries.
- Cloud based so on any machine is easy to share.

↳ Use cases:
- Busines performance tracking
- Sales analysis
- Financial reporting
- Operationdl Monitoring

Qlik

↳ - Data analytics & BI platform known for its associative data model
- Allows users to explore data w/o being limited by predefined queries.

↳ Key features:
- Fast in-memory processing.
- Connects w/ various data sources.
- Users can create their own dashboards & reports
- Enables free form data exploration.

↳ Use cases:
- Market & customer analysis
- Healthcare analytics
- Supply chain optimization
- Risk mgmt.

QlikView → Script-based & developer-driven
Qlik Sense → more modern, user-friendly, focused on self-service analytics

Apache Pig : (acts as bridge bet" user & Hadoop)

- Component of Apache Hadoop ecosyst".
- High level scripting language.
- Used for creating MapReduce programs used in Hadoop.
- Uses scripting lang. called **Pig latin** that simplifies coding req. for processing large datasets.
- Supports batch & interactive modes.

↳ Architecture :

① Pig Latin Scripts - Used for writing data analytics logic.

② Grunt Shell - Interactive C.I for users to write & execute Pig latin scripts.

③ Pig Server - Acts as interface bet" user & Pig runtime env., managing execution of Pig scripts.

④ Parser - Takes Pig latin script & converts into logical plan. I/P ↳, O/P ↳ (I/P)

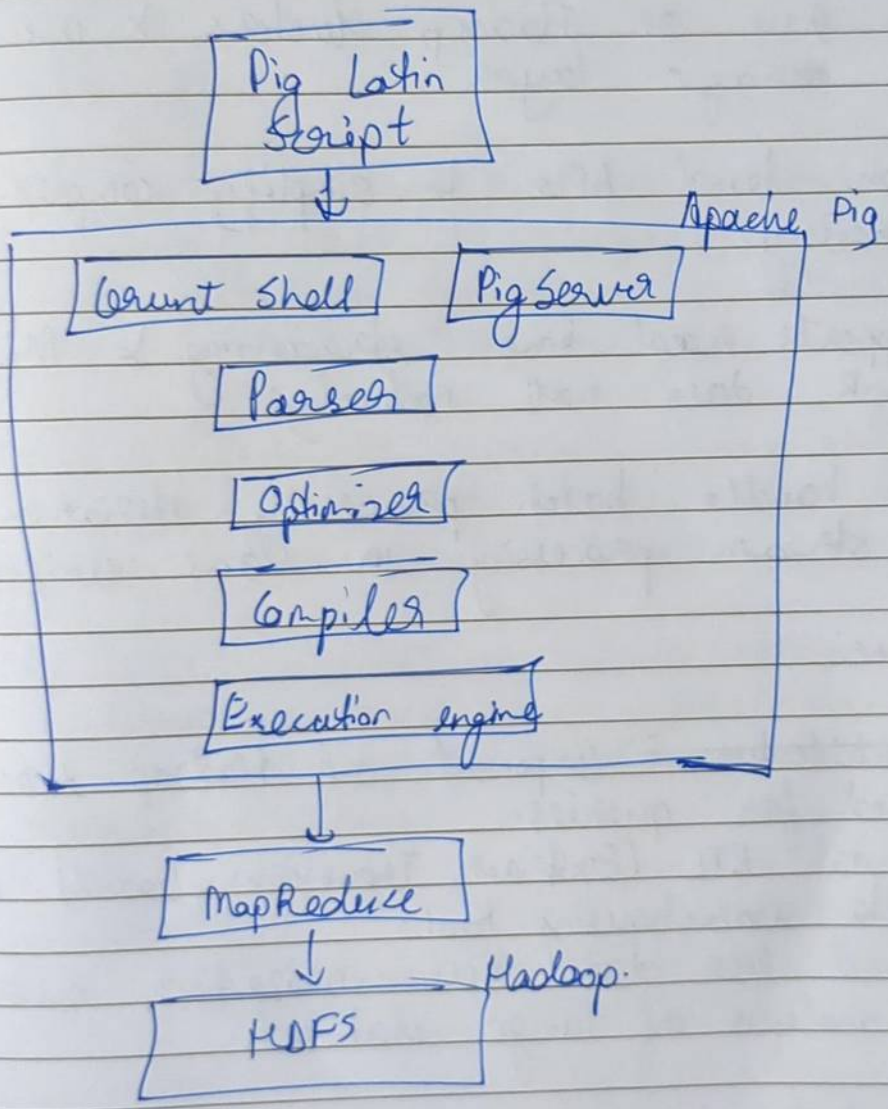
⑤ Optimizer - Improves logical plan by applying optimization rules to make data processing more efficient. O/P → optimize logical plan ↓

⑥ Compiler - Takes I/P ↳ & translates into series of MapReduce Jobs.

⑦ Execution engine - Runs compiled MapReduce

jobs on Hadoop cluster, processing data stored in HDFS.

⑧ MapReduce executes data processing tasks. HDFS stores I/P & O/P data.



↳ Role of Pig in Apache Hadoop:
- Pig Scripts converted to MapReduce Jobs auto. by Pig engine.
- Improves productivity of developer by abstracting complex MapR. logic.
- Useful for iterative data processing & prototyping.

Spark complementing Hadoop:

- Can run workloads upto 100× faster than traditional MapR job due to in-mem. data processing engine.
- Can run on Hadoop clusters & use HDFS as storage layer.
- High level APIs to simplify complex data processing.
- Supports real time streaming & ML which MapR does not natively.
- Can handle batch processing iterative algos & stream processing in one unified engine

Hive

- ~~Architecture~~ : Component of Hadoop ecosyst^n Used for queries.
- It is ETL (Extract, Transform, Load) & data warehousing tool.
- Used for data summarization, querying & analysis of large datasets.
- Architecture:

① UI - Hive has a CLI for writing & running HiveQL queries & a web UI to interact w/ Hive.

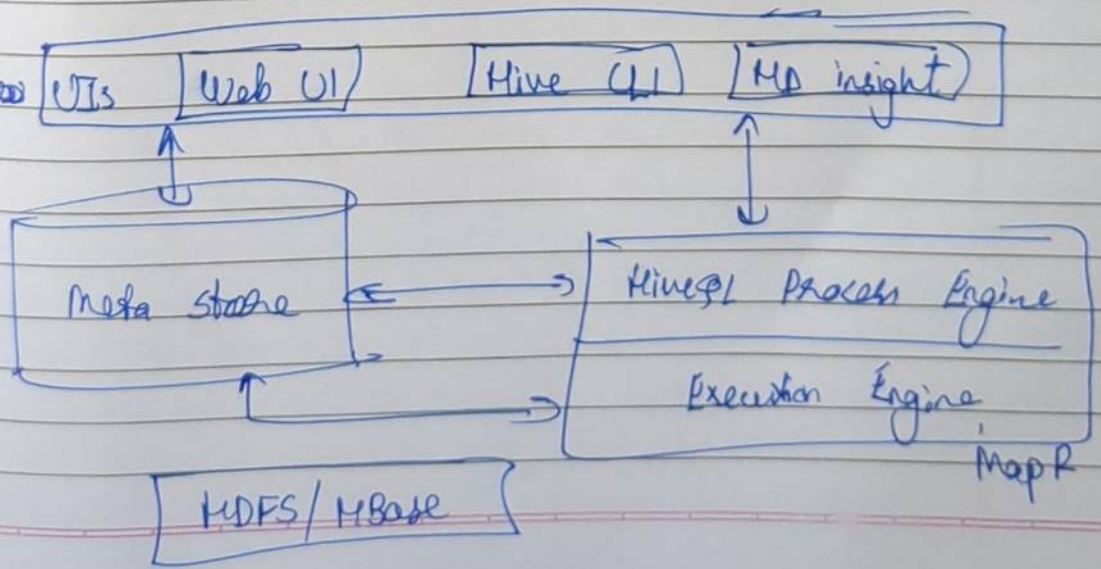
② Meta store – stores metadata about Hive tables (schema, data types, table location in HDFS). Managed using traditional RDBMS like MySQL.

③ HiveQL Process Engine – Accepts & processes HiveQL (SQL-like) queries. Parses, & compiles & optimizes the queries. Converts HiveQL queries to execution plans.

④ Execution engine – Works w/ process engine & MapReduce. Executes query plans by launching corresponding jobs & handles job monitoring & collects final result.

⑤ MapReduce – Default execution framework for Hive. HiveQL queries are internally converted to MapR. programs. Enables parallel processing.

⑥ HDFS / HBase – HDFS is default storage syst^{n} for Hive. HBase is used when real-time read/write access is needed. Stores data in tables in distributed format.



↳ Characteristics:**
- Has SQL like lang. (HiveQL) thus easy for users familiar w/ SQL.
- Schema is applied when reading data.
- Batch processing for large scale offline analysis
- Metadata stored in RDMS-based metastore

↳ Features:
- Scalable
- Compatible w/ Hadoop
- Supports summerization, querying & reporting
- Supports partitioning & bucketing to improve performance of data access.