Class Note:

Subject : Data analytics

Faculty : ....D. Jyothi.....

Topic: Types of analysis & key steps

Unit No. : 1
Lecture No.: 1
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 1

Types of Analysis and key steps:

- Descriphve analysis
- Exploratory Analysis
- Diagnostic Analysis
- predichve analysis
- prescriphve analysis

key steps in analysis

1. Identify the need
2. Collect data
3. clean data
4. perform data analysis
5. present the result.

Class Notes:

Subject : Data Analysis

Faculty : D. Jyothi

Topic: Types of Analysis

Unit No. : I
Lecture No.:
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 2

Types of Analysis.

1. Descriptive Analysis.

Descriptive analysis involves describing and presenting broader dataset in a way that allows emergent patterns to be observed from them to see if there are any obvious patterns. Data aggregation techniques are one way of performing descriptive analysis.

2. Exploratory analysis

Exploratory analysis involves consulting various data sets to see how certain variables may be related, or how certain patterns may be driving others. This analytic approach is crucial in framing potential hypotheses and research questions that can be investigated using data analytic techniques.

3. Diagnostic analysis

Diagnostic analysis used to answer why a particular pattern exists in the first place. This assist company

Diagnostic analytics includes methods such as hypothesis testing, determining a correlation vs causation, diagnostic regression analysis

4. predictive analysis.

predictive analysis answer the question of what happens. this type of analysis is key for companies in deciding new features or updates or enriching products, and in determining what product will perform well in the market.

5. prescriptive analysis

prescriptive analysis involves determining the most effective strategy for implementing the decision arrived at - for Eg. an organization can use prescriptive analysis to shift through the best way to unroll a new feature. this component of data analytics actively deals with the consumer end, requiring one to work with marketing, human resources and so on.

Class Note:

Subject : Data Analytics

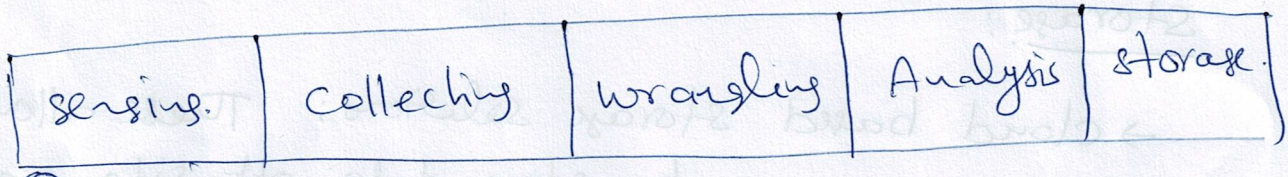
Faculty : D. Jyothi

Topic: Components of Modern Data Eco system.

Unit No. : I
Lecture No.: 8
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 4

Components of Modern Data Eco systems.

The term data ecosystem refers to the programming language, packages, algorithms, cloud-computing serices and general Infrastructure an organization uses to collect, store, analyze and leverage data.



① sensing.
Internal data sources and External data sources. software Algorithms

② collecting:
Programming Languages, code packages & APIs are used to extract the information

③ Data wrangling:
Is a set of processes designed to transform raw data into a more usable format. by using Algorithms, programming Languages, Data wrangling tools like. OpenRefine, Data wrangler, CSVkit.

Analysis

Depending on the specific challenges your data project seeks to address, analysis can be diagnostic, descriptive, predictive or prescriptive.

Datavisualization tools like Microsoft BI, google charts, which can create graphical representation of Data.

Storage:

→ cloud based storage solutions: These allow an organization to store data off-site and access it remotely.

→ On-site servers: Control over how data is stored and used.

→ other storage media: Eg: USB drives, CD-ROMs, floppy disks.

Class Note:

Subject : .....D.A.....

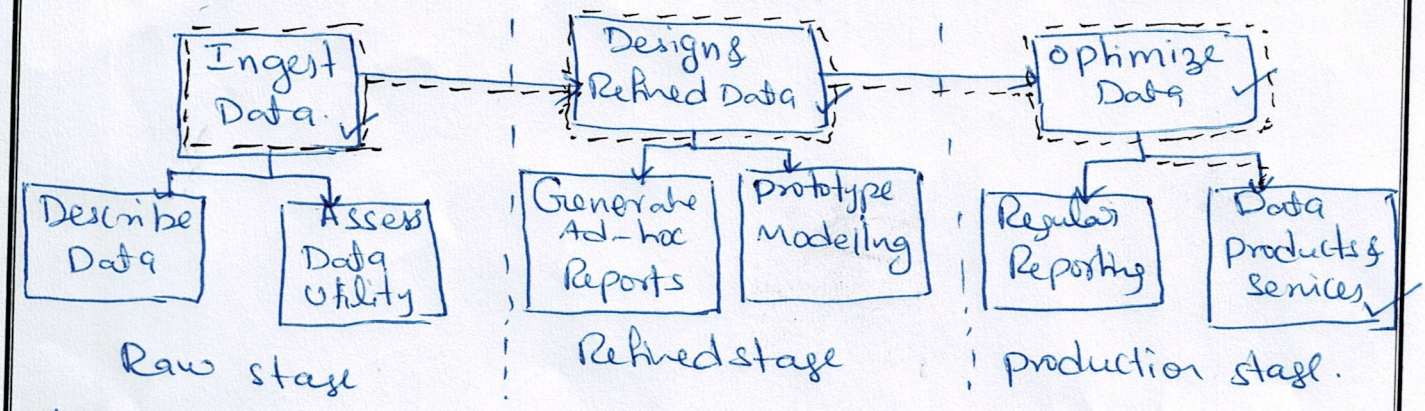
Faculty : .....D. Jyothi

Topic: ...Data Engineer

Unit No. : I
Lecture No.: 3
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:

Data Engineer.

Data Engineers are responsible for creation and upkeep of the systems that store, procen, and move data.



background knowledge needed. ----→ Flow of data wrangling framework.

- System administration
- Data processing algorithms & Implementation.

Data Architect

Data architects create catalogs for data to improve its discoverability and usability.

- optimizations
- Creation of naming conventions
- Standard documentation pratices.
- User requirement gathering procen of where to source the data and how to organize.
- Creating data schemas & alignment conventions.
- Major concentration on Design & Refine Data. in the Refined stage & optimize data on production stage.

Class Note:

Subject : .....DA.....

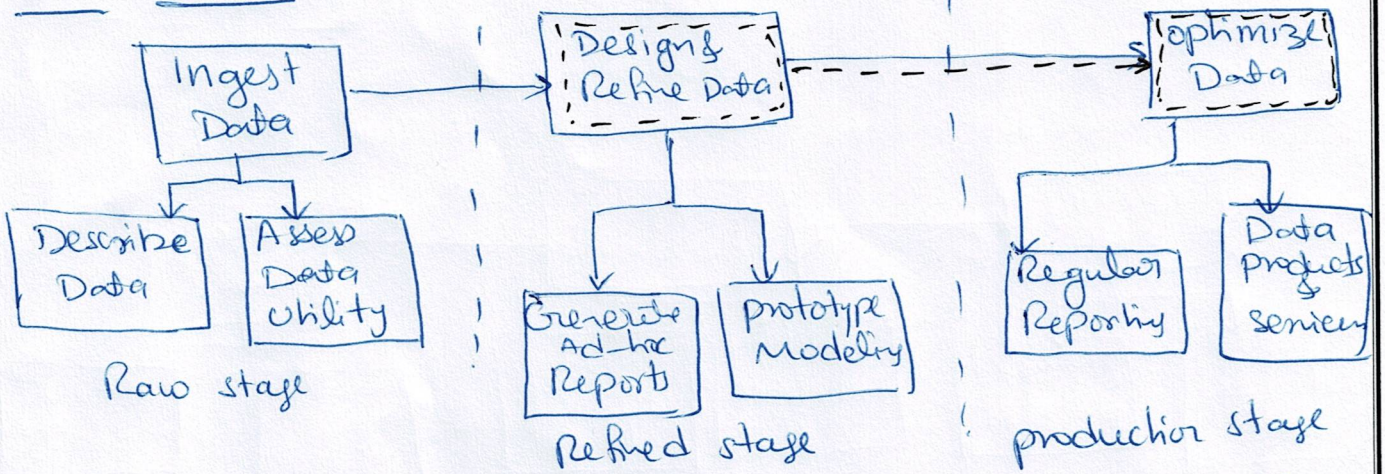
Faculty : .....D. Jyothi.....

Topic: Data scientist

Unit No. : I
Lecture No.: 3
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:

Page No. 7

Data Scientist:



- Data scientists are responsible for finding & verifying deep or complex sets of insights.
- Insights derived from existing data using advanced statistical analyses or from the application of machine learning algorithms.
- conducting experiments like modern A/B tests.
- productionalizing the insights.
- statistics focused works on A/B testing kind of methods and Engineering focused works on. prototyping & building data driven services & products.
- knowledge on mathematics & statistics algorithms
- Technologies used are R, SAS, Python, SPSS.

Class Note:

Subject : ..... D.A .....

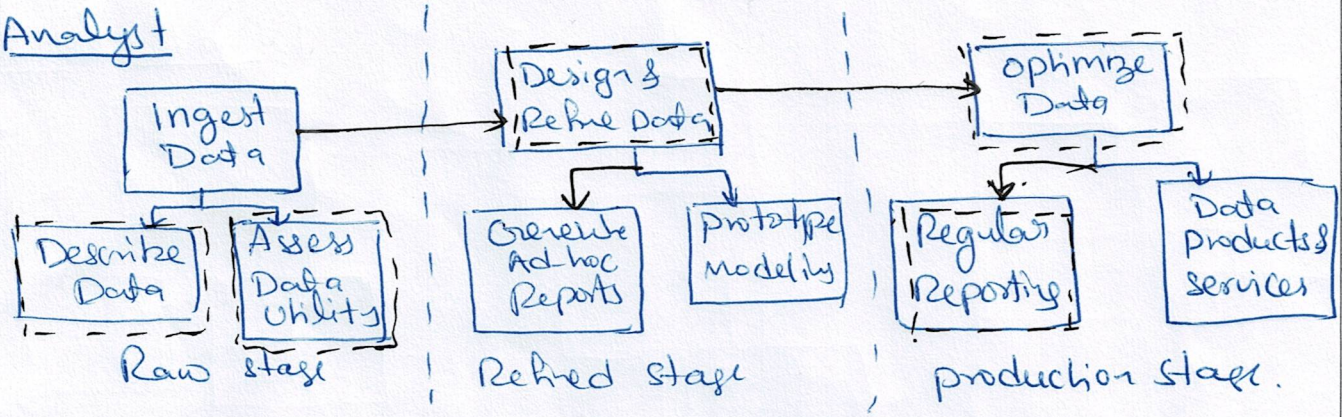
Faculty : ..... D. Jyothi

Topic: Data Analyst

Unit No. : I
Lecture No.: 4
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:

Page No. 8

Analyst



* Finding & delivering actionable insights from data is the major responsibility of the analyst.

- Analyst are responsible for providing a busines or organization with critical information.

- Identifying correlated indicators of KPI trends to better understand the underlying dynamics of the system

- strong knowledge on mathematics & statistics

* Analyst are strong system thinkers.

Class Note:

Subject : DA

Faculty : D. Jyothi

Topic: Types of Datastructures

Unit No. : 1
Lecture No.: 5
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 10

Data:
Anything that is recorded is data. observations & facts are data. Anecdotes and opinions are also data, of a different kind. Data can be numbers like the record of daily weather or daily sales. Data can be alphanumeric, such as the names of Employees.

Indirect value: Data provides value to your organization by influencing peoples decisions or inspiring changes in procen. Eg: risk modeling in insurance

Direct value: Data provides value to your organization by feeding automated systems. Eg: Netflix.

Raw	Refined	productis.
- Ingest Data - Data discovery & metadata creation.	- Create Canonical data for wide spread consumption • conduct analysis modelling and forecasting	• Create production quality data - Build regular reporting & automated data products / services.

Fig: Data moves through stages

KPI : key performance Indicator (KPI)

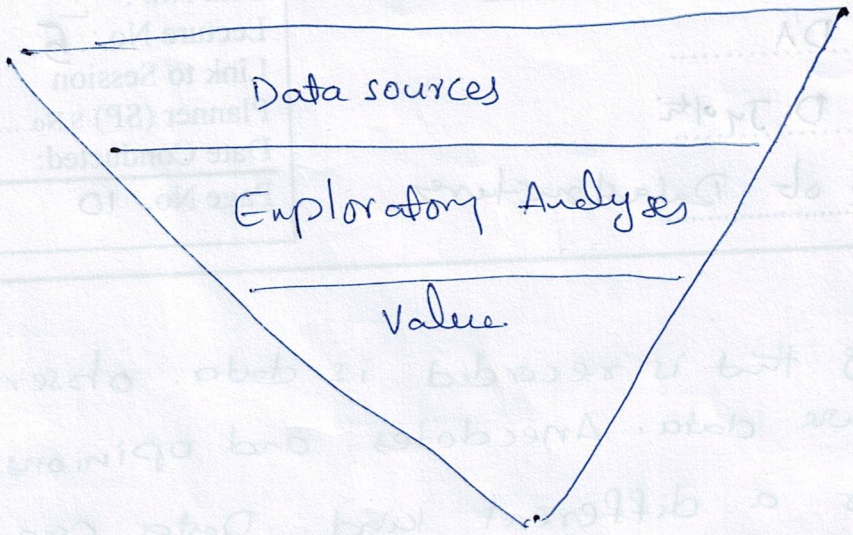


Fig: Data value funnel.

Data source: Empower the people who know your business priorities to explore your data.

Exploratory analytics: Effort should be as efficient as possible. this brings us back to data wrangling. the faster you can wrangle data, the more explorations of your data you can conduct, and the more analyses you will be able to move into production

Value:
Efficient data wrangling workflow can enable more business analysts to explore a larger quality of data at a faster pace.

Class Note:

Subject : ....DA.....

Faculty : ....D.Jyothi....

Topic: .....

Unit No. : I
Lecture No.:
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 12

Types of Data structures

The structure of a dataset refer to the format and encoding of records and fields.

Tool Data structures

ExcelGrid
SQLTabular (uniform records)
Trifacta wrangler versions.

Excel:

Excel required data to be laid out in a grid and grid does not need to be rectangular or completely filled. people include multiple tables in a single Excel grid.

- Mix descriptive text with data,
- Embed graphics within their spreadsheets.
- not strictly rectangular or consistant
- Each cell of a grid supports a wide variety of value types. from numbers and percentage to dates and times.

SQL:

- SQL expects datasets to be constructed as a set of records, in which every record contains the same set of fields.
- wrangle using SQL must be rectangle & specific schema.
- Different versions of SQL support different fields types Eg: basic set of dates, times, strings and numbers and universal.

Trifacta wrangler:

- handles structured, unstructured and unstructured data.
- Data does not need to be explicitly broken down into rows & columns or fully populated.
- Trifacta supports a variety of different data types from basic integers to more complex custom types like dates, US states, phone numbers.

SQL:

- SQL expects datasets to be constructed as a set of records, in which every record contains the same set of fields.
- wrangle using SQL must be rectangle & specific schema.
- Different versions of SQL support different fields types Eg: basic set of dates, times, strings and numbers and universal.

Trifacta wrangler:

- handles structured, unstructured and unstructured data.
- Data does not need to be explicitly broken down into rows & columns or fully populated.
- Trifacta supports a variety of different data types from basic integers to more complex custom types like dates, US states, phone numbers.

Class Note:

Subject : .....DA.....

Faculty : .....D. Jyoti.....

Topic: .....File format.....

Unit No. : I
Lecture No.: 6
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:

File Formats:

- CSV (Comma-separated values): Data is organized in rows & columns, with each value separated by commas.
- JSON (JavaScript object notation): JSON is a lightweight and human readable data interchange format. It is used for structured data representation and is commonly used in web applications & APIs.
- XML (Extensible Markup Language): XML is another markup language used for structuring and storing data in a hierarchical format. It is used for data interchange & configuration files.
- Parquet: Columnar storage file format
- Avro: Avro is a data serialization system that provides rich data structures and a compact binary format.
- ORC (Optimized row columnar): ORC is columnar storage file. It offers lightweight compression and improved performance for read-intensive workloads.

Apache Arrow

Apache Arrow is a cross language development platform for in memory data. It provides a standard for representing data in memory.

Excel: (XLSx format) used for tabular data storage, business and financial data.

HDF5 (Hierarchical Data format).

HDF5 is a flexible file format for storing and managing large volumes of scientific data. supports complex data structures.

SQLite: SQLite is a lightweight, serverless database system that uses a single self-contained file to store data. Used for small-scale applications and embedded systems.

Class Notes:

Subject : ....D.A.....

Faculty : .....D..J.y.o.t.s.....

Topic: .....

Unit No. : I
Lecture No.:
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 16

File formats

Common file formats used in Data science

- CSV
- Text file
- JSON
- Microsoft Excel file
- SAS
- SQL
- Python pickle file
- Stata
- HDFS
- HTML
- ZIP
- PDF
- DOCX
- Images
- Google Bigquery

CSV

CSV stands for Comma Separated values which is a text based file format that store data in a tabular form similar to a spreadsheet.

```
import pandas as pd
filename = "C:\\Text file.csv"
data = pd.read_csv(filename)
```

Reading from Microsoft Excel file.

The important file difference between XLS and XLSX is XLS is a binary file XLSX is a open XML format.

```
import pandas as pd
filename = "C:\\TextFile.xlsx"
df = pd.read_excel(name, sheetname = "Test")
```

Reading a ZIP file:

ZIP is a archieve file format that supports lossless data compression. You can read a ZIP file by importing the "zipfile" module.

```
import zipfile
archive = zipfile.ZipFile('Test.zip', 'r')
df = archive.read('Test.csv')
```

Reading data from SQL

SQL stands for structured Query Language SQL lets you managing data held in a relational database management system. Here the code.

Class Notes:

Subject : .....DA.....

Faculty : .....D...Jyothi

Topic: .....

Unit No. : I
Lecture No.:
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 18

```
import pandas as pd
from sqlalchemy import create_engine
    engine = create_engine('sqlite://Master.sqlite')
    with engines.connect() as con:
    rs = con.execute('SELECT * from orders')
    data = pd.DataFrame(rs.fetchmany(size=5))
    data.columns = rs.keys()
```

Reading data from pickle file:

Its file type native to python and these files are serialized i.e convert the object to bytestream. The pickle module implements binary protocols for serializing and deserializing python object structure.

```
        import pickle
filename = "C:\\testfile.pkl"
    with open(filename, 'rb') as file:
        data = pickle.load(file)
    print(data)
```

Class Note:

Subject : DA

Faculty : D. Jyothi

Topic : Data sources.

Unit No. :
Lecture No.: 7
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 19

Sources of Data

1. Operations Data. Data from all busines applications, including from ERPs systems that form the backbone of an organization's IT systems. Data to be extracted will depend upon the subject matter of the data warehouse. Eg: customer care, marketing.

2. Specialized applications:

Applications such as (POS) point of sale terminals and e-commerce applications that also provide customer-facing data. supplier data could come from supply chain Management Systems. planning and budget data should also be added as needed for making comparisons against target.

3. External syndicated data:

This includes publicly available data such as weather or economic activity data. provides contextual information to decision makers.

Class Notes:

Subject : ..... DA .....

Faculty : ..... D. Jyothi

Topic: Data profesional Languages

Unit No. :
Lecture No.: 8
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 20.

Data profesional Laraguages.

1. python
2. SQL
3. R
4. Julia
5. Java script
6. Scala
7. Java
8. Go
9. MATLAB
10. C/C++
11. Statistical Analytical system (SAS)

Class Notes:

Subject : ..... DA .....

Faculty : ..... D. Jyothi .....

Topic: Various Data Repositories

Unit No. :
Lecture No.: 9
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 21

Various Data Repositories

- Google Dataset search: search engine for researchers to locate online data.
- datasetlist: offers a list of biggest ML datasets from across the word.
- UCI: Data classified by types of problems, attribute types, the field of study etc
- fastai-datasets - dataset for Image classification. NLP and image localization.
- NLP-datasets: Alphabetical list of free/public domain datasets with text data for use in NLP
- Bifrost: For visual classified by task, application class, label, and format.
- ImageNet
- CTMedical Images
- Flickr-faces.

Class Note:

Subject : DA

Faculty : D. Jyoti

Topic: ETL process

Unit No. : II
Lecture No.: 10
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 22

ETL process:

Extract:

- Data is extracted from diverse data sources, which can include databases, datawarehouse, API, logs, flat files, web services, and more.
- Data extraction method can vary depending on the source and may involve direct querying, using APIs, bulk data exports, or real-time data streaming.
- The extracted data is often in its raw and original form, including structured, semi structured, and unstructured data.

Transform:

- The extracted data undergoes a series of data transformations to ensure its quality, consistency, and compatibility with the target data repository.
- Data cleaning & validation: Data is cleaned to remove duplicates, fix errors, handle missing values, and resolve inconsistencies.

Data Integration:

Data from multiple sources is combined and integrated into a unified format, resolving differences in data structures and schema.

Data enrichment:

Additional data may be derived or added to enrich the dataset, such as calculating derived metrics, merging data from different sources, or translating data into a common language.

Data aggregation:

Data may be aggregated to create summary datasets or to group data at different levels of efficient analysis.

Load

- The transformed data is loaded into the target data repository, which can be a data warehouse, a data lake, a database, or any other data storage system
- Data is stored in a way that makes it accenible and optimized for analysis and reporting.
- Loading strategies can differ based on the size of the data & target repository

Class Note:

Subject : DA

Faculty : D. Jyothi

Topic: .....

Unit No. :
Lecture No.: 11
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:
Page No. 24.

Introduction to Big data:

Big data refers to extremely large & complex datasets that exceed the processing capabilities of traditional data management systems. These datasets are characterized by their volume, velocity, and variety 3 Vs of big data.

Volume: Big data involves massive amount of data that are too large to be handled by conventional database systems. It includes data generated from various sources, such as social media, web interactions, sensors, transactions, and more.

Velocity: Big data is generated and collected at high speed and often in real time or near real time. The continuous and rapid flow of data requires efficient data processing & storage.

Variety: Big data encompasses different types of data, including structured, semi-structured, and unstructured data. It includes text, images, audio, video, logfiles, geospatial data and more.

Class Notes:

Subject : DA

Faculty : D. Jyothi

Topic: Big data Ecosystem

Unit No. :
Lecture No.: 12
Link to Session
Planner (SP) S.No. .....of SP
Date Conducted:

Big data Ecosystem:

The layers of Ecosystem are

- Data Ingestion
- Data storage
- Resource Management
- Data processing and analysis
- Data Access

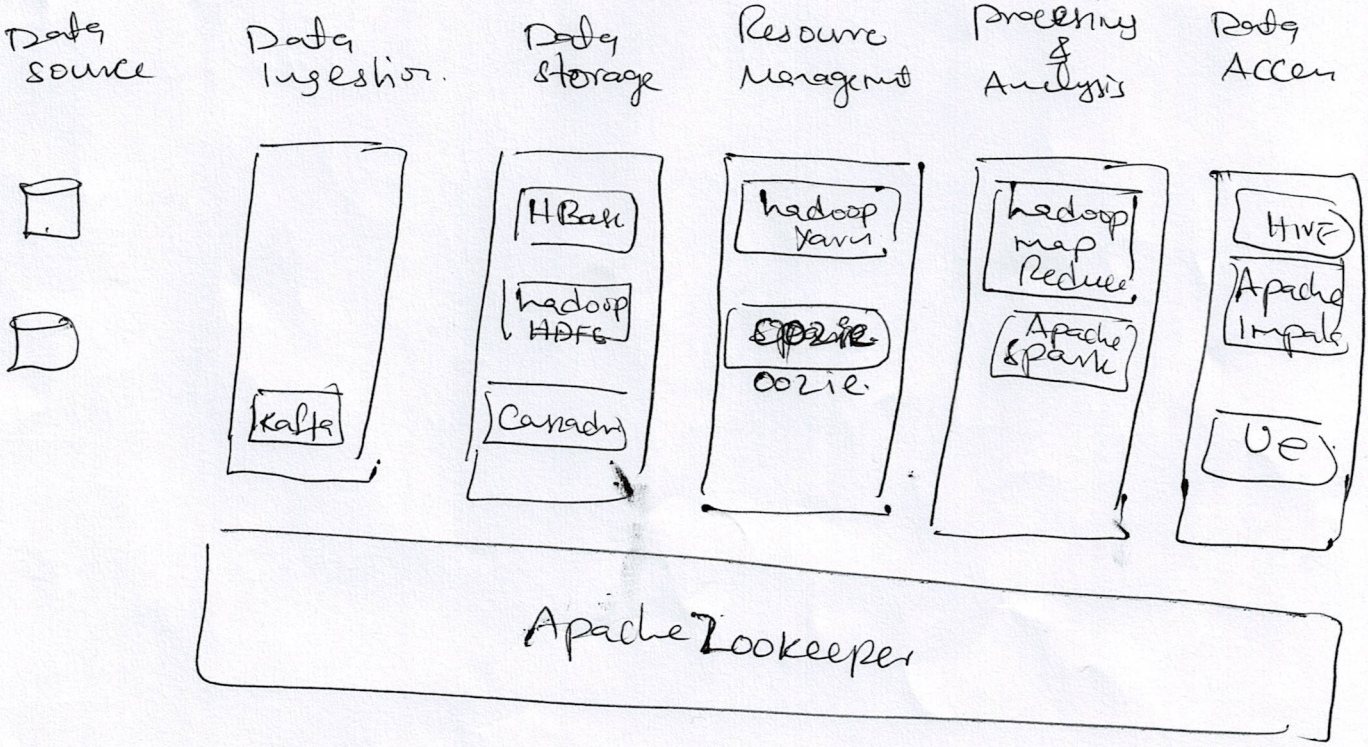


Fig: Hadoop Ecosystem