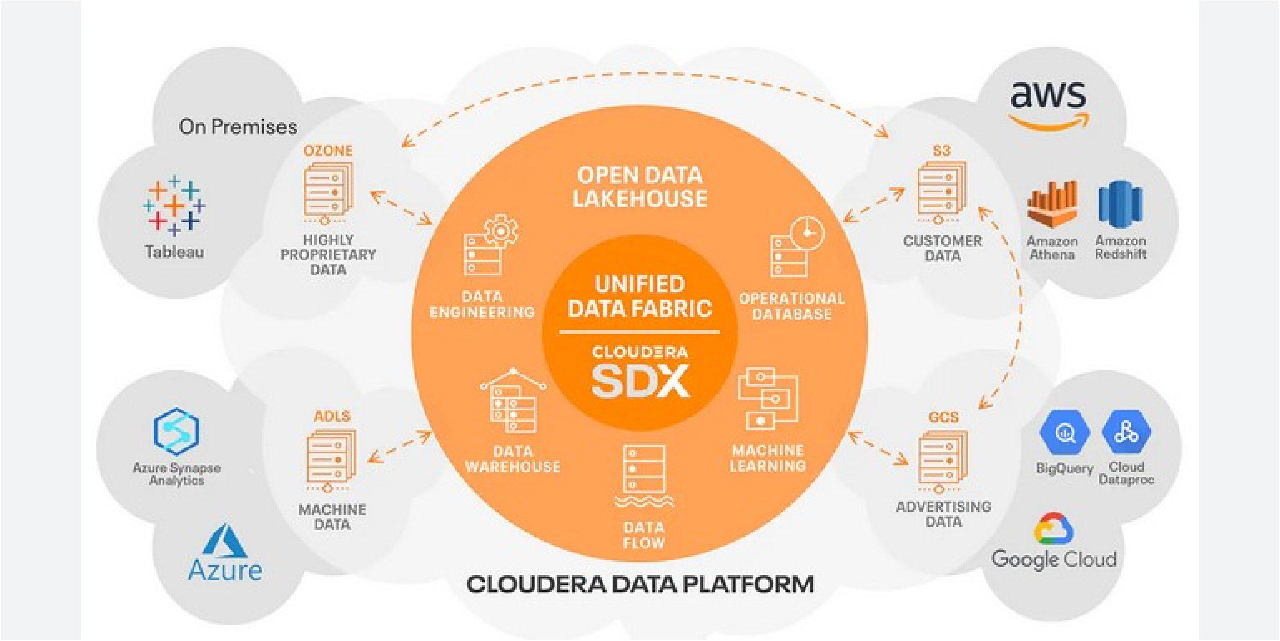
Latest Features on CDP :
1. Spark 3.1 with Python 3– Multi-versions support . Support for migration from Spark 2 to Spark 3
2. YoniKorn – Cloud Native Resources Scheduling on Kubernetes (K8s , https://kubernetes.io/)
3. AirFlow - Authors, schedules and monitors workflows| Load (ETL) workflow orchestration tool

Designing Data Lake with Cloudera CDP – all questions comes here!!!
1. Generic Architecture:



Open Data Lakehouse Architecture

Latest Technology available from Cloudera for Data Lake House:
1. Iceberg : It is a high-performance format for huge analytic tables.
2. YoniKorn : Cloud Native Resources Scheduling on Kubernetes
3. AirFlow : Authors, schedules and monitors workflows| Load (ETL) workflow orchestration tool

2. Cloudera CDP and Technologies

3. NoSQL Database – capable of supporting 3-5 TB Database
1. IBM Cloudant
2. RavenDB, 3. CouchBase
3. As *Cloudera's* OpDB and Apache HBase

https://www.predictiveanalyticstoday.com/top-nosql-document-databases/

4. Resources Scheduling – YoniKorn

5. ETL/Workflow Management/Orchestration

6. System Security- ISO 27001 Compliance (Default)

7. Outdated BI Of Past, Automated AI of Future Paradigm
Next Data Architecture= Data Mess+ Data Fabric+ Data Lake House
Data Mess -

Data Febric – Standard
Best Analytics – Tableau (https://www.tableau.com/)

CDP Public Cloud : No Installation, Cloudera SAS
Create and manage secure data lakes, self-service analytics, and machine learning services without installing and managing the data platform software. CDP Public Cloud services are managed by Cloudera, but unlike other public cloud services, your data will always remain under your control in your VPC. CDP runs on AWS, Azure and Google Cloud
https://www.cloudera.com/products/cloudera-data-platform.html?tab=0

CDP Public Cloud Data Services:

Data Flow
Data Processing
Data Engineering
Data Warehouse
Operational Database
Machine Learning
Data Hub

CDP Private Cloud: Installation on Cloud/Data Center, private installation
It delivers powerful analytic, transactional, and machine learning workloads in a hybrid data platform. With a choice of traditional as well as elastic analytics and scalable object storage, CDP Private Cloud modernizes traditional monolithic cluster deployments in a powerful and efficient platform
https://www.cloudera.com/products/cloudera-data-platform.html?tab=1

CDP Private Cloud Data Services:

Data Engineering
Data Warehouse
Machine Learning

Cloudera SDX: Cloudera SDX delivers an integrated set of security and governance technologies built on metadata and delivers persistent context across all analytics as well as public and private clouds.

Accessing CDP P

1. Running Reports on CDP Data Analytics: - Base Cloudera HDP/valid for Cloudera CDP Data Analytics too
1. Login to Data Analytics Interface, http://ip:30800
2. Check Database and tables for necessary datasets/data
3. Write or place Query for Report, SELECT Account_ID, Store_Cost, unit_Sales from account
4. Execute Query and analyse the report

Solve Data Problems using HDFS:

HDFS Commands, Sqoop Data Movement (https://sqoop.apache.org/ retired),
NiFi Data Flow, importing data in HBase, Impala

Solve Data Problems using Hive :

Solve Data Problems using Impala :

Solving Data Processing using Spark :

Apache NiFi and Jenkis on Ubuntu VM:

Moving data from RDBMS to HDFS/Hive **- Solutions**

1. Sqoop – retired

2. Spark -

```
from pyspark import SparkContext
sc = SparkContext(appName='SparkWordCount Example')
print("Go and give 5 times", sc)
inputFile = sc.textFile('C:\\InputFile.txt')
print(inputFile)
counts = inputFile.flatMap(lambda line: line.split()).map(lambda word: (word, 1)).reduceByKey(lambda a, b: a + b)
print("Counted:", counts)
#counts.saveAsTextFile('C:\\Results\Output.txt')
sc.stop()
```

Querying Insights/Data from Data Lake House:

1. Hive Virtual Warehouse with Tableau :

https://docs.cloudera.com/data-warehouse/1.2/querying-data/topics/dw-connect-to-hive-virt-warehouse-with-tableau.html

1. Get Tableau Desktop from https://www.tableau.com/
2.

2. Hive Virtual Warehouse with RStudio :

R-Studio: https://posit.co/