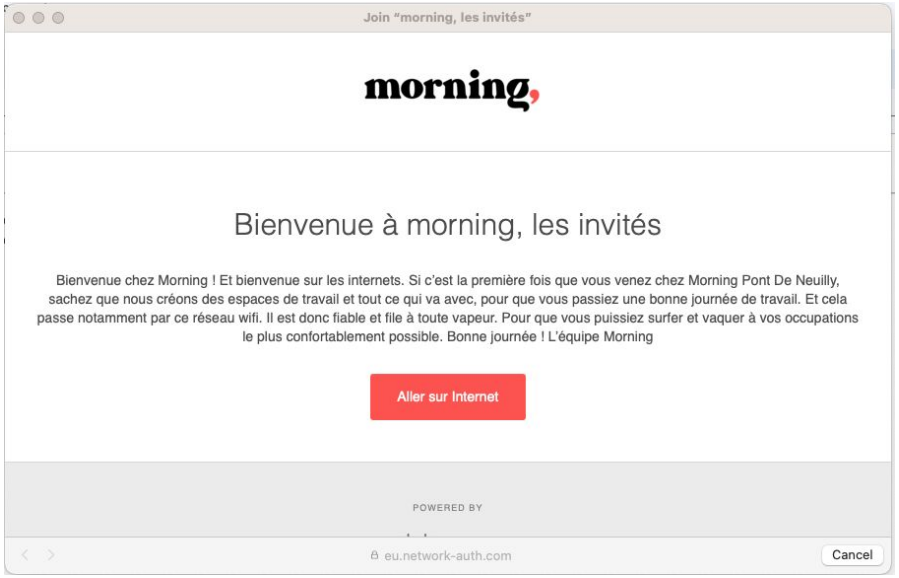
Les Ateliers

CLOUDERA

Data Lifecycle in the Public CLoud

LOGISTICS

Wifi



Other

- Toilettes
- Break
- Questions/answers

AGENDA

- 08:30 Accueil & Petit Déjeuner
- 09:00 Introduction
- 09:30 Data Ingestion
- 10:00 Data Engineering
- 10:30 Coffee Break
- 10:45 Machine Learning
- 11:15 Data Warehouse
- 11:45 Q&A et Labs Optionnels
- 12:20 Clôture et debrief

INTRO DUCTION

WHO YOU ARE

NAME
COMPANY
ROLE

Users

user001	David DRUBIGNY
user002	Laurent BAULAY
user003	Khalil Cherif
user004	Yonni Hervé
user005	EL MOOTAZ LAMAA
user006	Benoît Hagenbourger
user007	Cedric TOUZET
user008	Hamza Khribi
user009	Thierry GUERIN

user010	Dingan LIAO
user011	priti Bista
user012	Khalil Gharbi
user013	Matthieu BARRET
user014	Stéphane Andreu
user015	Ibrahima Matar Gueye
user016	

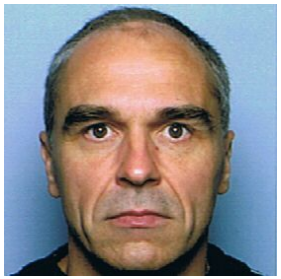
WHO WE ARE

CLOUDERA TEAM

Jacques Marchand
Solutions Engineer
jmarchand@cloudera.com

Patrick Cousin
Solutions Engineer
pcousin@cloudera.com

Charles Aad
Solutions Engineer
charles.aad@cloudera.com

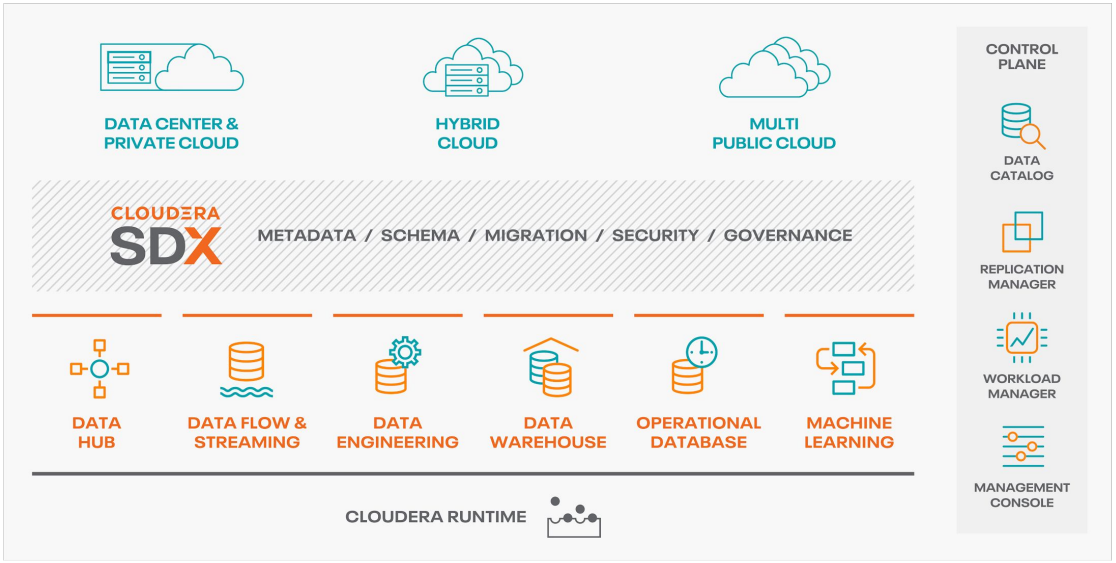


Olivier Meignan
Solutions Engineer
olivier.meignan@cloudera.com

CLOUDERA PUBLIC CLOUD

CLOUDERA DATA PLATFORM

- View one pane of glass across **hybrid** and multi-clouds
- Scale to petabytes of data and 1,000s of diverse users
- Control cloud costs with auto scale, suspend and resume
- Optimize workloads based on analytics and machine learning
- Inspect data lineage across clouds and clusters

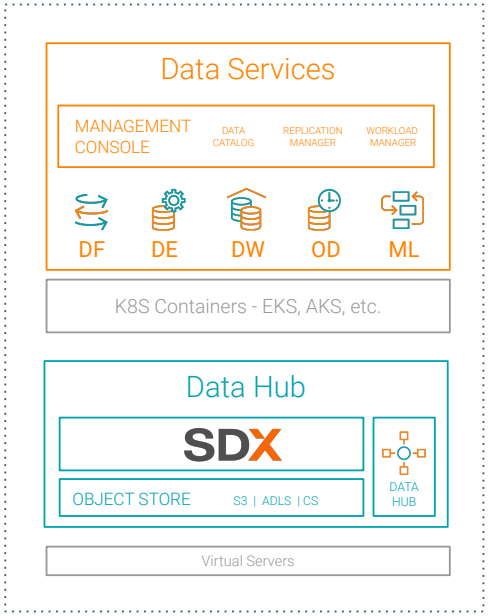


CDP PUBLIC CLOUD

Cloud-native architecture with containerized Experiences and Base cluster foundation

Admin and user experience is consistent across CDP Private & CDP Public for true hybrid cloud

CDP Public Cloud



Analytic Experiences

- Ideal for new apps, bursty workloads and demanding users
- Easier to operate, 10x faster
- Auto scale, suspend and resume
- Better analytics user experience, outperforms shadow IT

Data Hub

- Storage management and SDX for Analytic Experiences
- Supports existing apps and traditional workloads with Data Hub
- Data Hub offers cluster management for Flow, Streaming, Data Engineering, Data Warehouse & Operational Database

KEY TECHNOLOGIES FOR THE DATA LIFECYCLE

Machine Learning

Collaborative ML workspaces for data scientists to develop, experiment and deploy models into production with secure, self-service enterprise data access

Data Engineering

Schedule, monitor, and debug data pipelines to streamline ETL processes quickly and securely with built-in job scheduling and troubleshooting

Data Visualization

Curate fast, self-service dashboards, reports and charts to easily and quickly develop and share agile analytical insight across your business

Data Hub

Easily manage data clusters across the data lifecycle running Apache Spark, Hive, Impala, HBase, Phoenix, NiFi, Kafka, Flink, and more

DataFlow & Streaming

Scalable, real-time streaming platform to ingest, curate, and analyze data with an easy no-code approach to developing sophisticated streaming applications easily

Data Warehouse

Deploy easy-to-use data warehouses with high performance SQL engines for teams of business analysts that need sub-second query response times on petabyte scale data

Operational Database

High-performance NoSQL database with unparalleled scale and performance for business critical operational applications

Shared Data Experience

Security, governance and metadata technologies that reduce security risks and operational costs through central policy controls that are automatically enforced across analytics in public and private clouds

BENEFITS OF CDP PUBLIC CLOUD

Simplify Data Analytics

You Own Your Data

First Class Security

Hybrid Flexibility

Common Skill Set

Data Lifecycle

Easy and Portable

WORKSHOP ENVIRONMENT

Users

user001	David DRUBIGNY
user002	Laurent BAULAY
user003	Khalil Cherif
user004	Yonni Hervé
user005	EL MOOTAZ LAMAA
user006	Benoît Hagenbourger
user007	Cedric TOUZET
user008	Hamza Khribi
user009	Thierry GUERIN

user010	Dingan LIAO
user011	priti Bista
user012	Khalil Gharbi
user013	Matthieu BARRET
user014	Stéphane Andreu
user015	Ibrahima Matar Gueye
user016	

CONNECTING TO LAB ENV

https://login.cdpworkshops.cloudera.com/auth/realms/field-marketing-emea/protocol/saml/clients/cdp-sso
user0XX (your account, do not forget it)
G0yvxvdms5srhyKF

Step 1: change your Workload Password in the Management Console (click on your user name/profile)
"Paris2024"

Repo:
https://github.com/charlesaad/Paris-Atelier-Cloud-24

Data Services


DataFlow


Data Engineering


Data Warehouse


Operational Database


Machine Learning

Data Management


Data Hub Clusters


Data Catalog


Replication Manager


Workload Manager

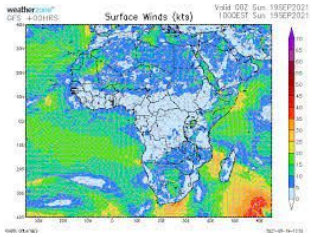

Management Console

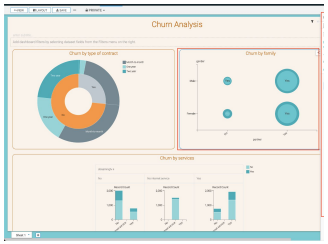
LAB INTRODUCTION

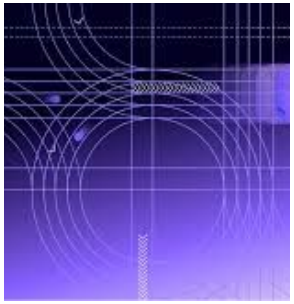
CLOUDERA

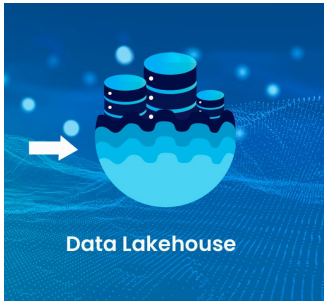
3000 Feet Use Case Description

Customer Churn

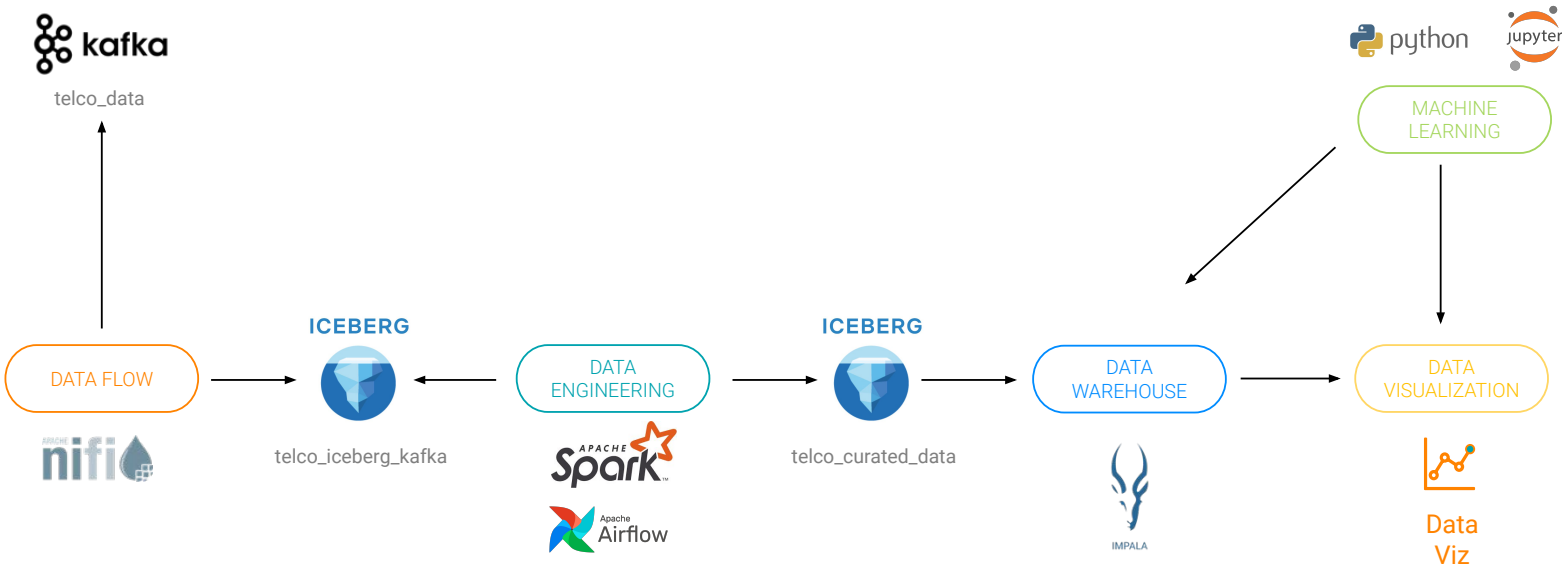




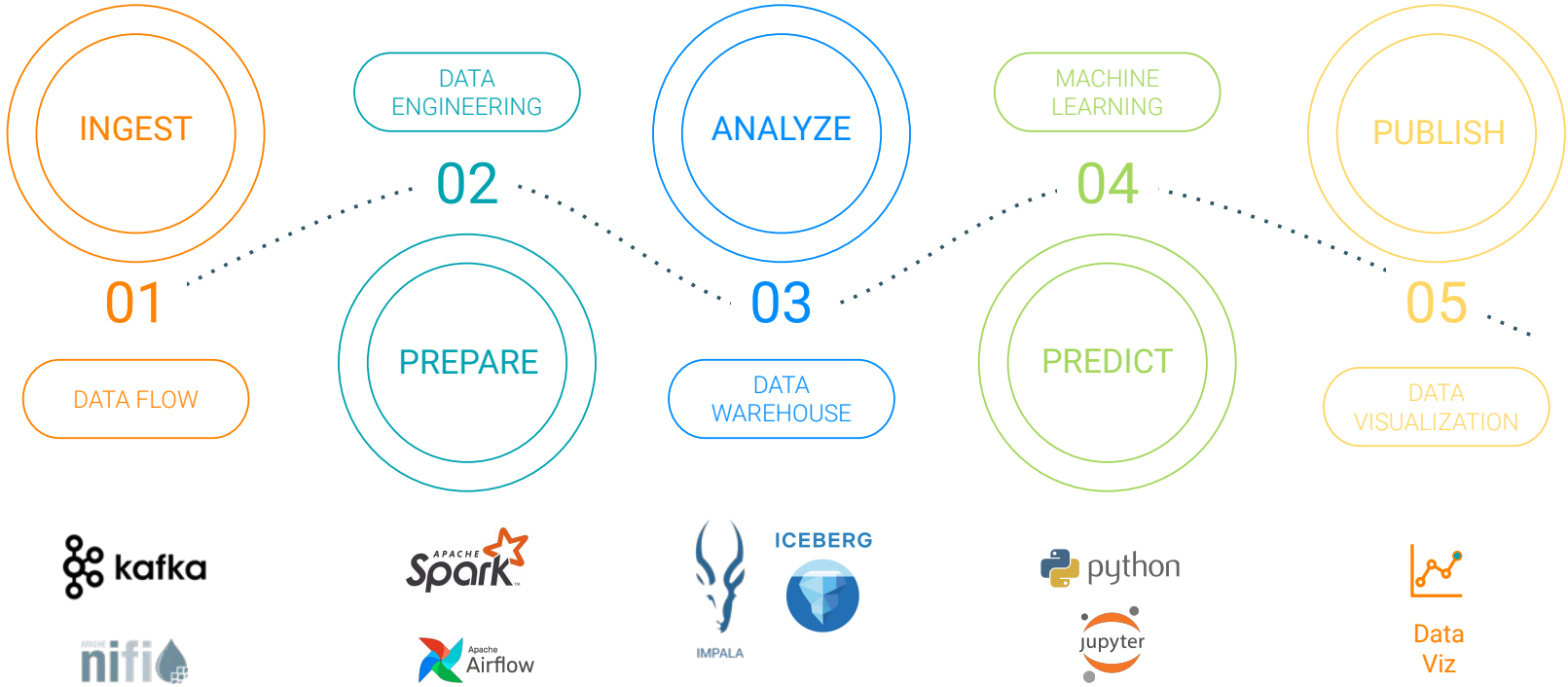




ANALYTICS ACROSS THE DATA LIFECYCLE



ANALYTICS ACROSS THE DATA LIFECYCLE



TODAY'S EXPECTATIONS

Data Lifecycle

1. Ingest real time data to an Open Lakehouse
2. Run Data Engineering transformation
3. Query/explore data and build Data Viz applications
4. Train and deploy a ML model to predict customer churn
5. Enrich data analytics with real time scoring

DATA SOURCE

customerID	7590-VHVEG	5575-GNVDE	3668-QPYBK	7795-CFOCW	9237-HQITU	9305-CDSKC
gender *	F	M	M	M	F	F
SeniorCitizen	0	0	0	0	0	0
Partner *	Y	N	N	N	N	N
Dependents *	0	0	1	0	0	0
tenure	1	34	2	45	2	8
PhoneService	No	Yes	Yes	No	Yes	Yes
MultipleLines	No phone service	No	No	No phone service	No	Yes
InternetService	DSL	DSL	DSL	DSL	Fiber optic	Fiber optic
OnlineSecurity	No	Yes	Yes	Yes	No	No
OnlineBackup	Yes	No	Yes	No	No	No
DeviceProtection	No	Yes	No	Yes	No	Yes
TechSupport	No	No	No	Yes	No	No
StreamingTV	No	No	No	No	No	Yes
StreamingMovies	No	No	No	No	No	Yes
Contract *	1	2	1	0	1	1
PaperlessBilling	Yes	No	Yes	No	Yes	Yes
PaymentMethod	Electronic check	Mailed check	Mailed check	Bank transfer (automatic)	Electronic check	Electronic check
MonthlyCharges	29,85	56,95	53,85	42,3	70,7	99,65
TotalCharges	29,85	1889,5	108,15	1840,75	151,65	820,5
Churn	No	No	Yes	No	Yes	Yes

CLOUDERA

* Attributes to be enriched

Are Cloudera Public Cloud Data Services

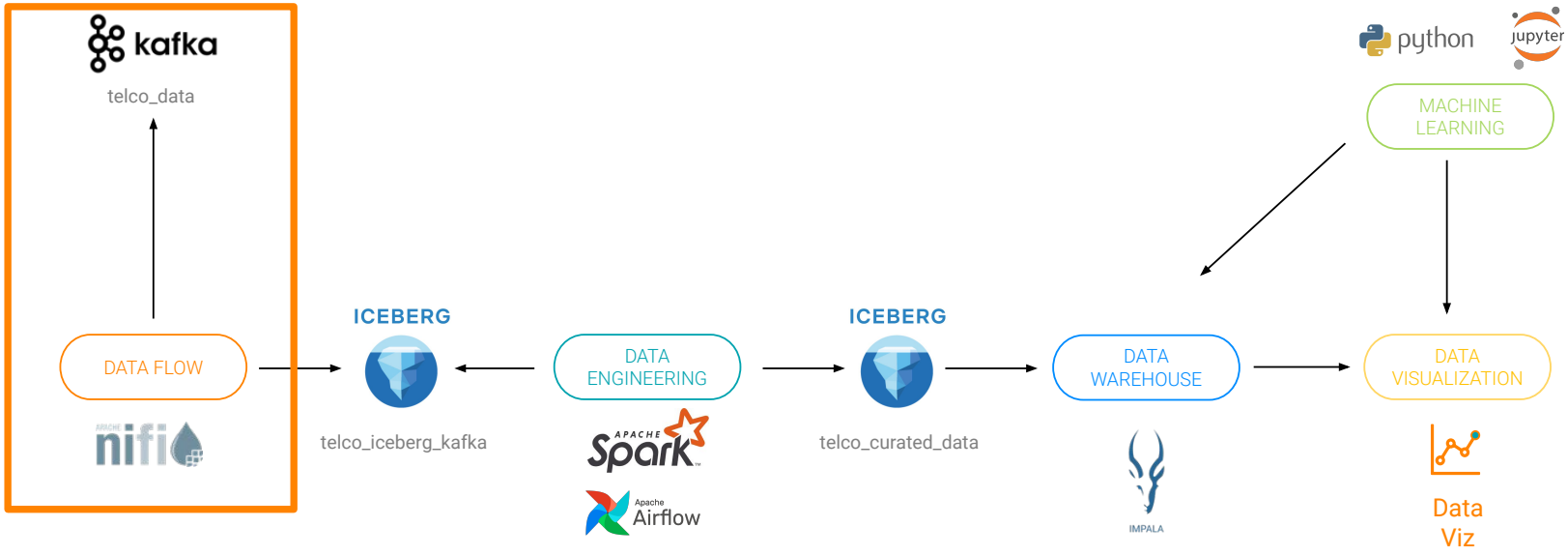
Choose one of the following

1. IaaS
2. PaaS
3. SaaS
4. DaaS
5. All of the Above
6. None of the above

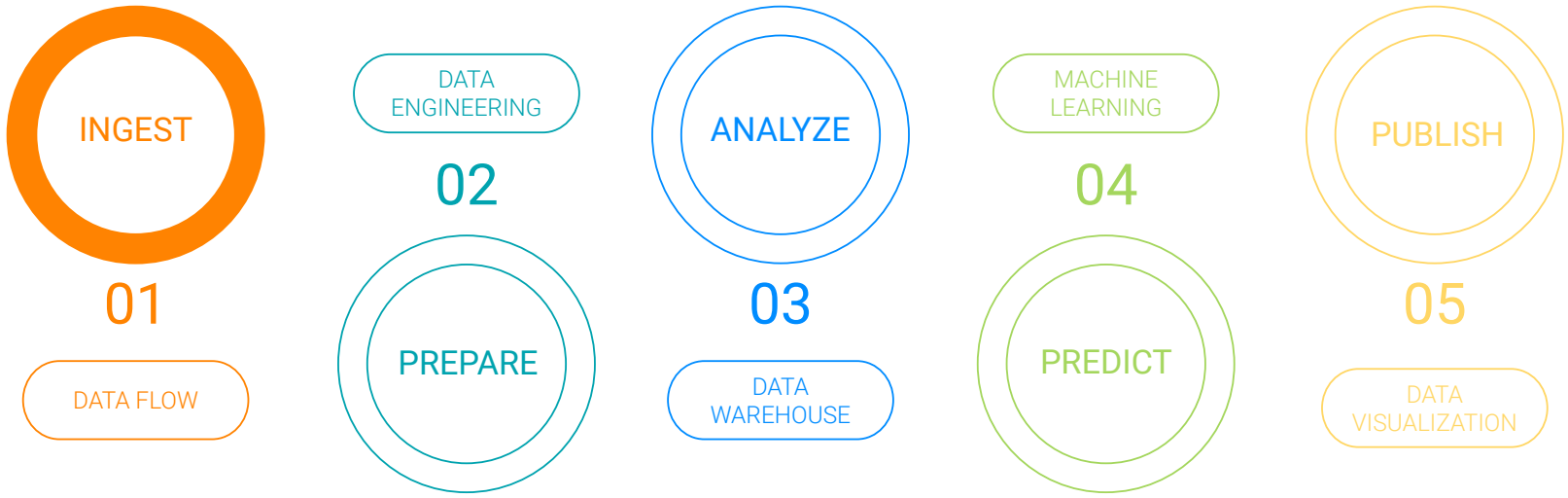
LAB 1:
Data Flow

CLOUDERA

ANALYTICS ACROSS THE DATA LIFECYCLE



ANALYTICS ACROSS THE DATA LIFECYCLE



CLOUDERA SDX

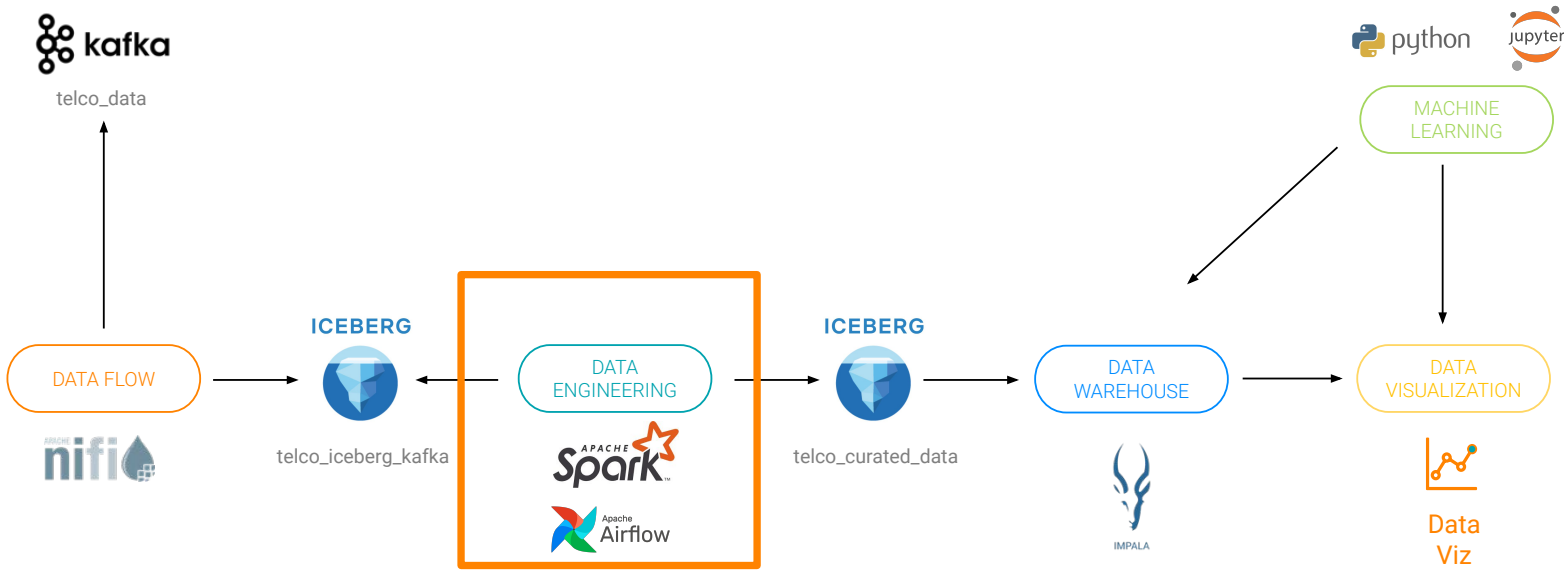
SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

LAB 1 – OVERVIEW

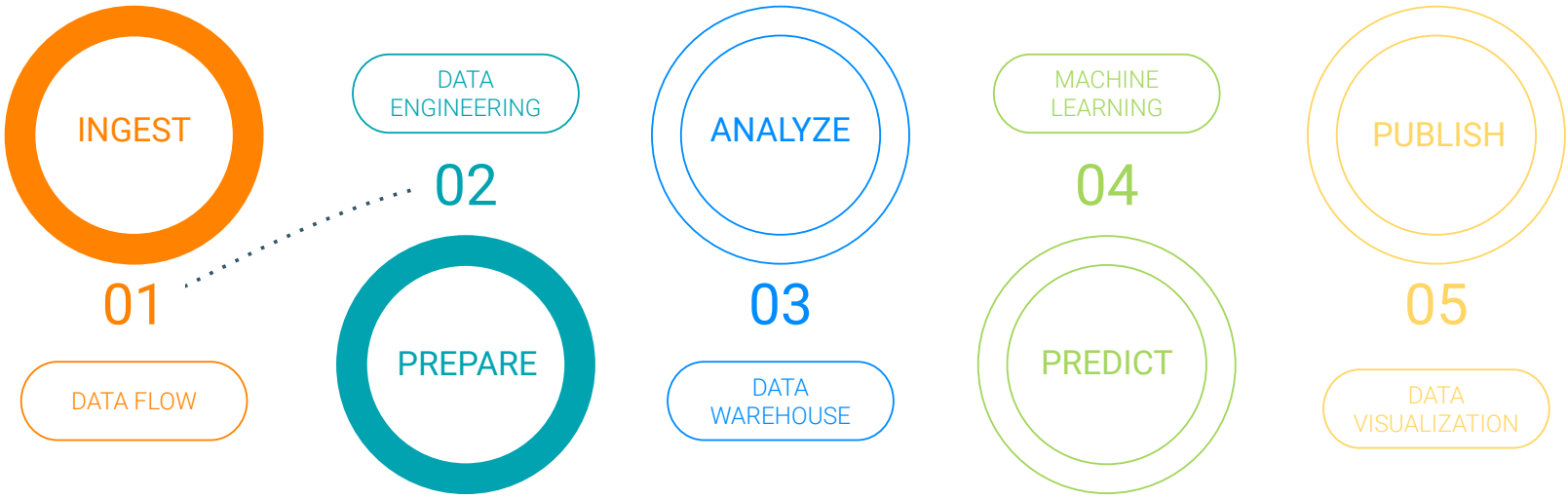
- CDF tour (5min)
- Configure and deploy a NiFi Flow from Catalog, to ingest data from Kafka topic to Open Lakehouse
- Execute the Flow pipeline

LAB 2:
Data Engineering

ANALYTICS ACROSS THE DATA LIFECYCLE



ANALYTICS ACROSS THE DATA LIFECYCLE



CLOUDERA SDX

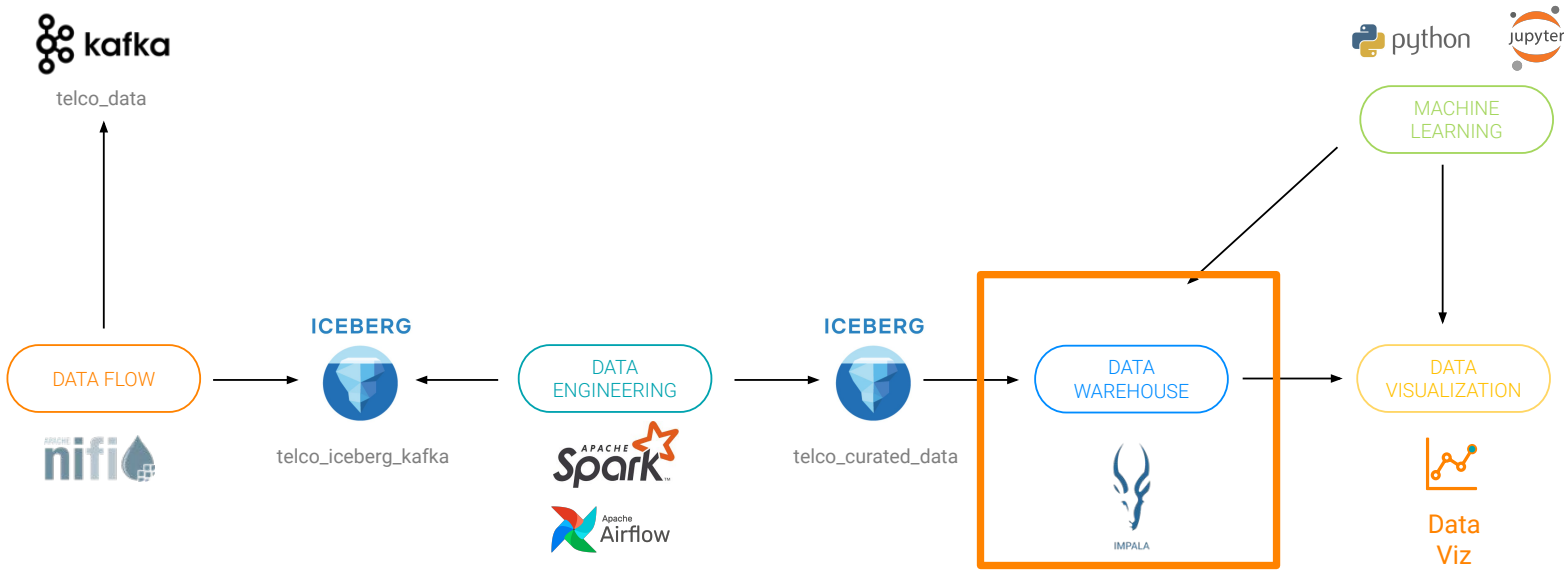
SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

LAB 2 – OVERVIEW

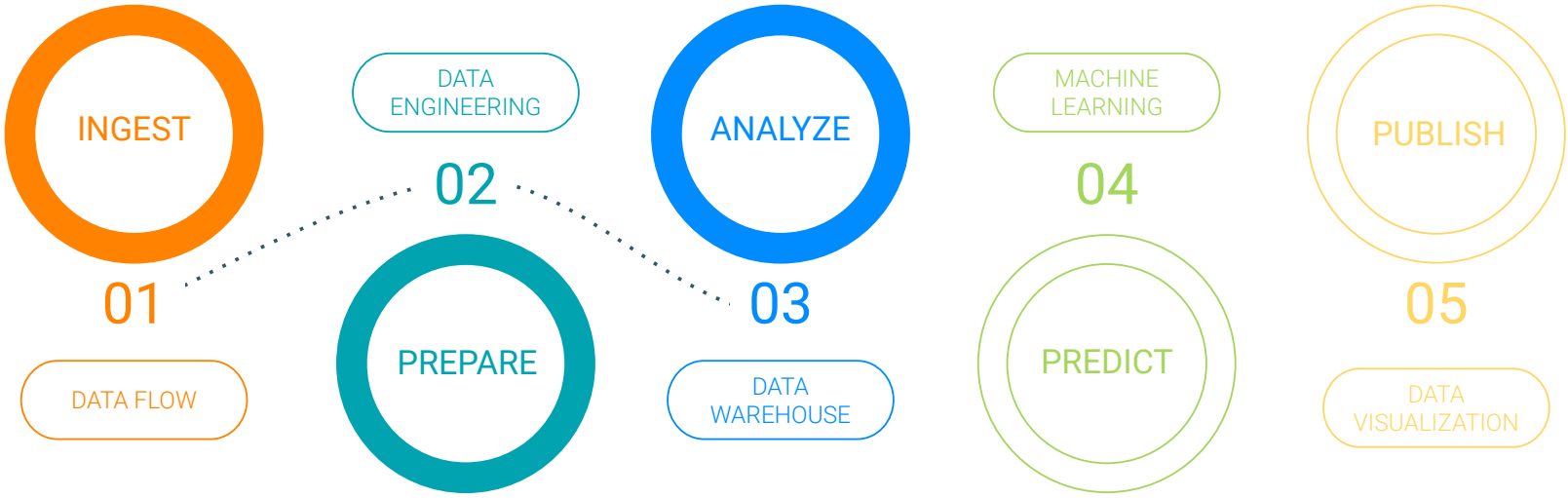
- CDE tour (5min)
- Configure and deploy Spark jobs with Editor
- Airflow to orchestrate the jobs
- Run data enrichment

LAB 3:
Data Warehouse

ANALYTICS ACROSS THE DATA LIFECYCLE



ANALYTICS ACROSS THE DATA LIFECYCLE



CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

LAB 3 – OVERVIEW

- CDW tour
- Query data from Open Lakehouse
- Build a dashboard

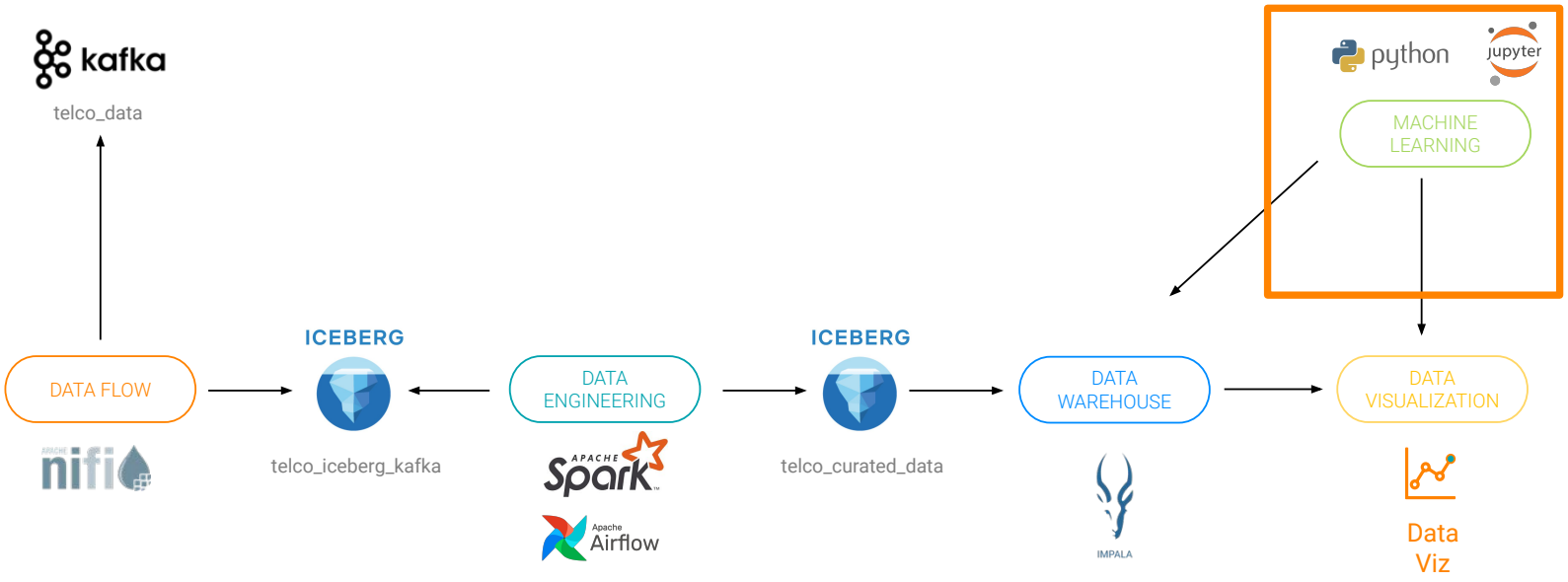
LAB 4:
Machine Learning

Which of the following is not a Cloudera Service

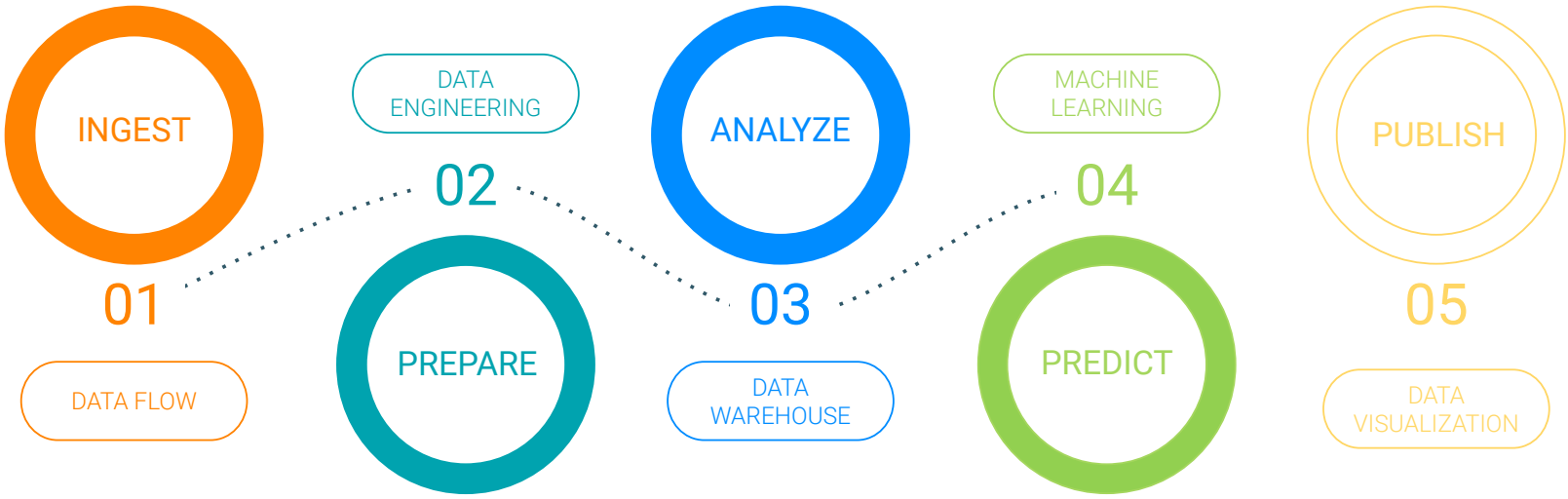
Choose one of the following

1. Data Flow
2. Data Engineering
3. Data Warehouse
4. Data Visualization
5. Operational Database
6. Machine Learning

ANALYTICS ACROSS THE DATA LIFECYCLE



ANALYTICS ACROSS THE DATA LIFECYCLE



CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

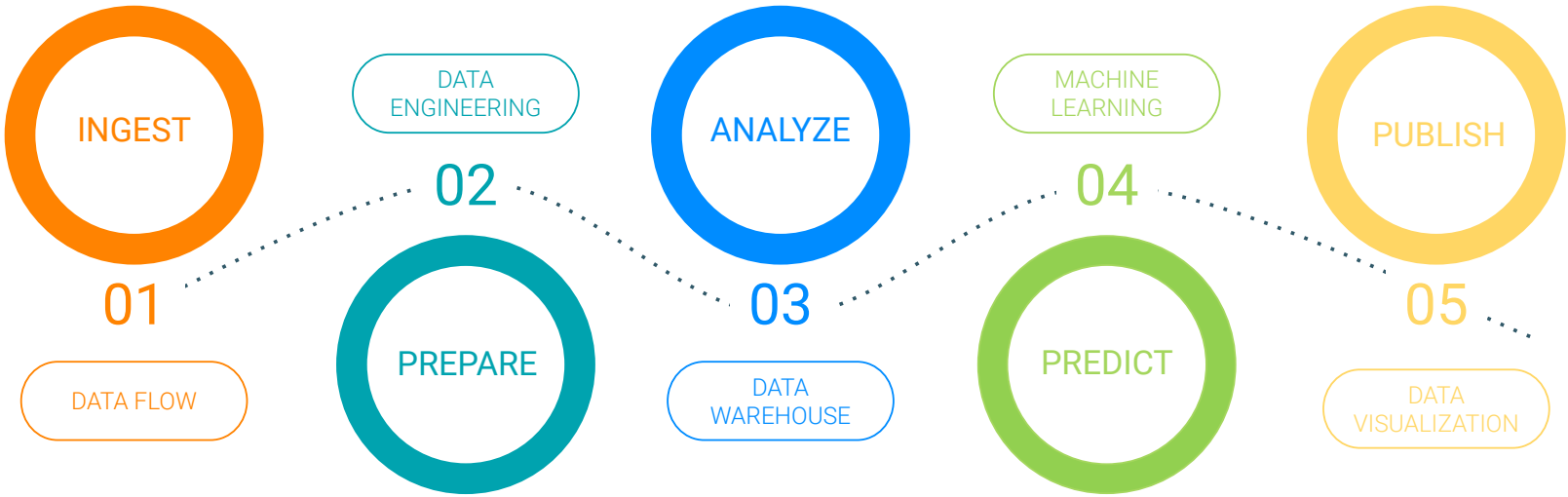
LAB 4 – OVERVIEW

- CML tour
- Train a ML model to predict customer churn
- Deploy the trained model for real time scoring/prediction

LAB 5:
Optional

CLOUDERA

ANALYTICS ACROSS THE DATA LIFECYCLE



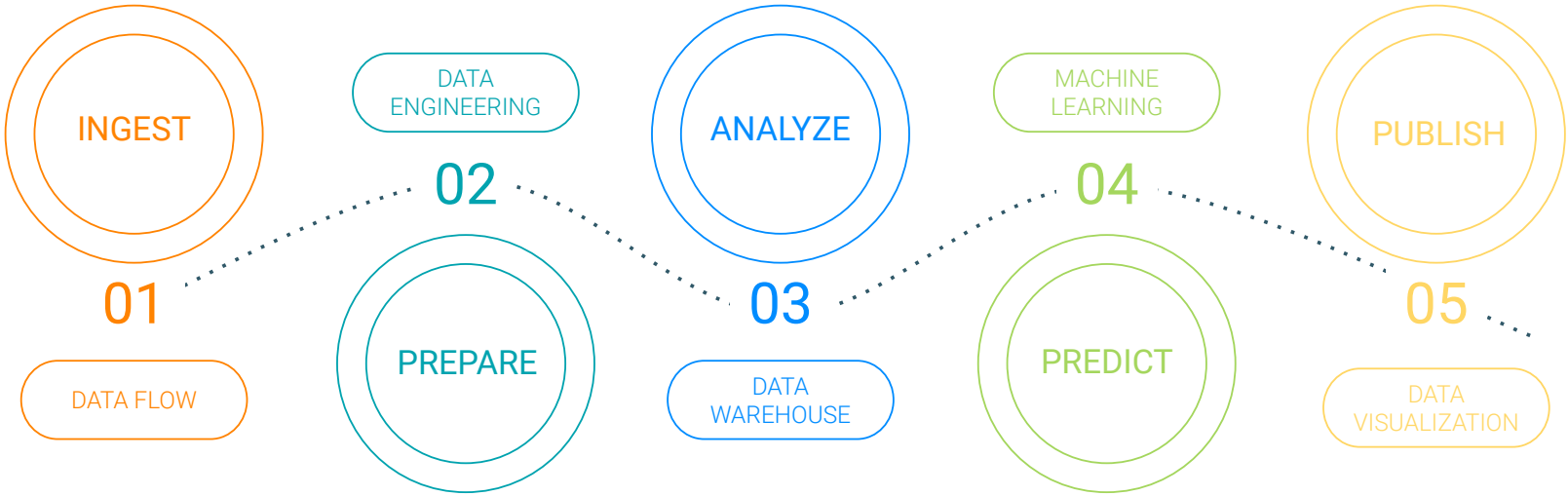
CLOUDERA SDX SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

LAB 5 – OVERVIEW

- Build customer 360 visuals with prediction model
- Connect to Data using Hue
- Iceberg
- Enrich Data Viz application with real time model scoring

WRAP UP

ANALYTICS ACROSS THE DATA LIFECYCLE



CLOUDERA SDX

SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

Feedback

IoT

CLOUDERA

CLOUDERA & IOT
REAL-TIME MONITORING OF YOUR HEART'S HEALTH. THANKS DATA.

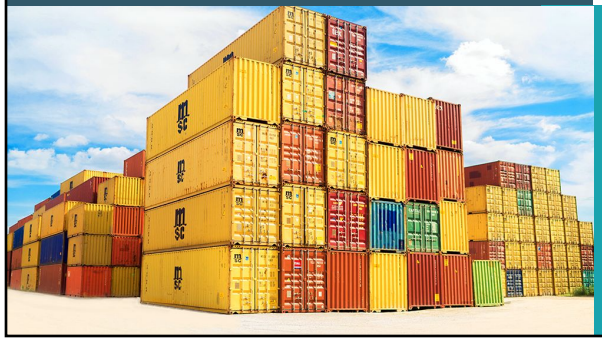
ENTERPRISE DATA CLOUD

CLOUDERA

CLOUDERA Now

EDGE 2AI

CLOUDERA



CLOUDERA

SDX

CLOUDERA

CLOUDERA HANDS-ON EXPERIENCE

Thank you!