

CLOUDERA

CLOUDERA

HANDS-ON

EXPERIENCE

Data Lifecycle

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	
user002	
user003	
user004	
user005	
user006	
user007	
user008	
user009	
user010	

user011	
user012	
user013	
user014	
user015	
user016	
user017	
user018	
user019	
user020	

WHO WE ARE

CLOUDERA TEAM



Jacques Marchand

Solutions Engineer

 jmarchand@cloudera.com



Cristina Sánchez

Solutions Engineer

 cristina.sanchez@cloudera.com



Charles Aad

Solutions Engineer

 charles.aad@cloudera.com

Why this hands on lab?

- 18 octobre - Paris présentiel
 - Atelier de modernisation de workloads dans le cloud public
 - <https://rb.gy/k172s>
- 19 octobre - Webinaire virtuel
 - Comment traiter vos flux en temps réel?
 - <https://rb.gy/rwi66>
- 9 novembre - Webinaire virtuel
 - Open Data Lakehouse
 - <https://rb.gy/va04f>
- 16 Novembre - Paris présentiel
 - Atelier flux de données
 - <https://rb.gy/bj4ip>

LOGISTICS

Wifi



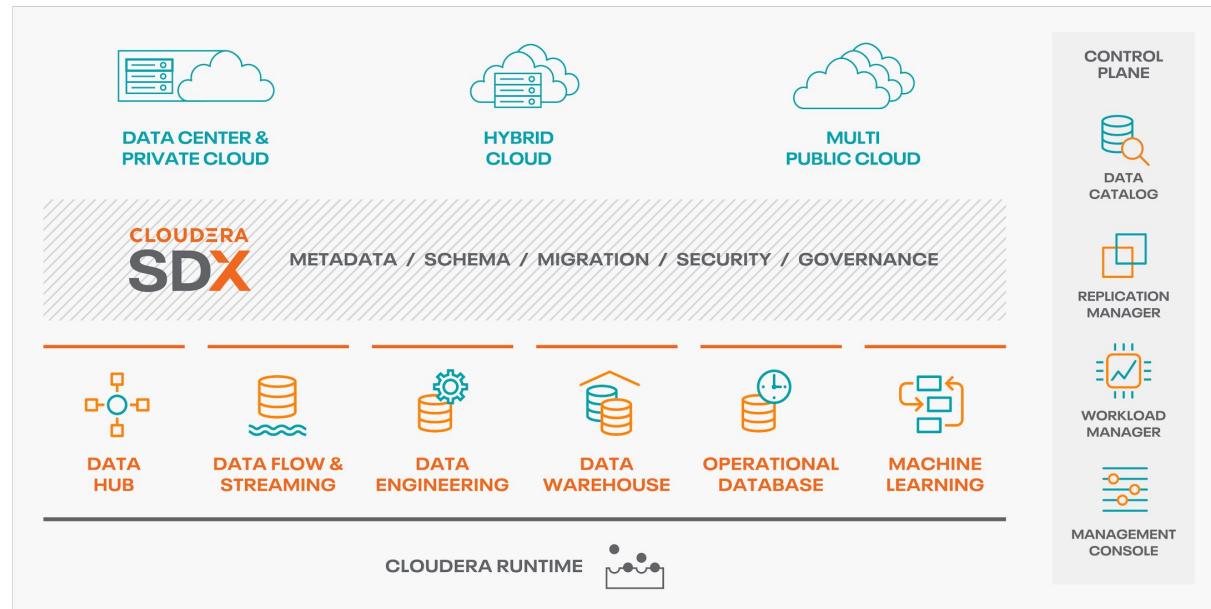
Other

- Toilettes
- Break
- Questions/answers

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

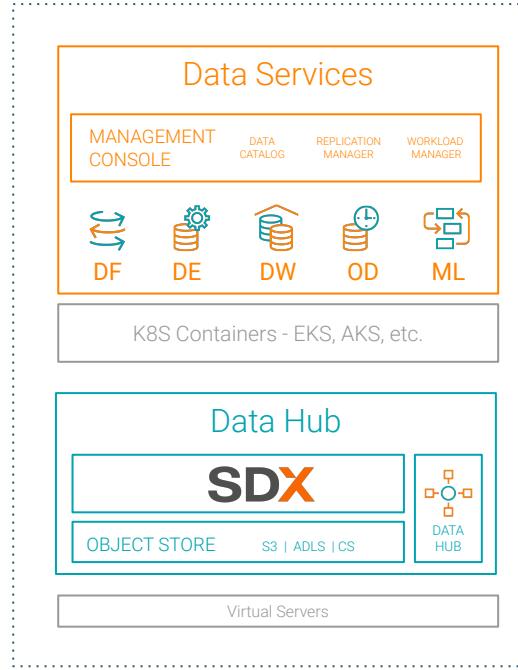


CDP PUBLIC CLOUD

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



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 Engineering

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



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



Data Visualization

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



Operational Database

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



Data Hub

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



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



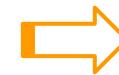
Hybrid
Flexibility



Common
Skill Set



Data
Lifecycle



Easy and
Portable

WORKSHOP ENVIRON MENT

Users

user001	
user002	
user003	
user004	
user005	
user006	
user007	
user008	
user009	
user010	

user011	
user012	
user013	
user014	
user015	
user016	
user017	
user018	
user019	
user020	

Data Services



Data Flow



Data Engineering



Data Warehouse



Operational Database



Machine Learning

Data Management



Data Hub Clusters



Data Catalog



Replication Manager



Workload Manager



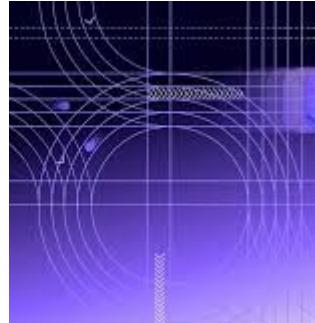
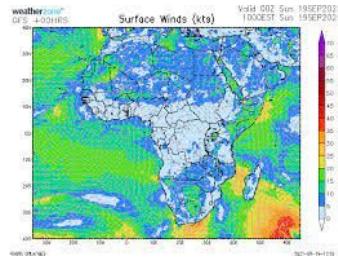
Management Console

LAB INTROD UCTION

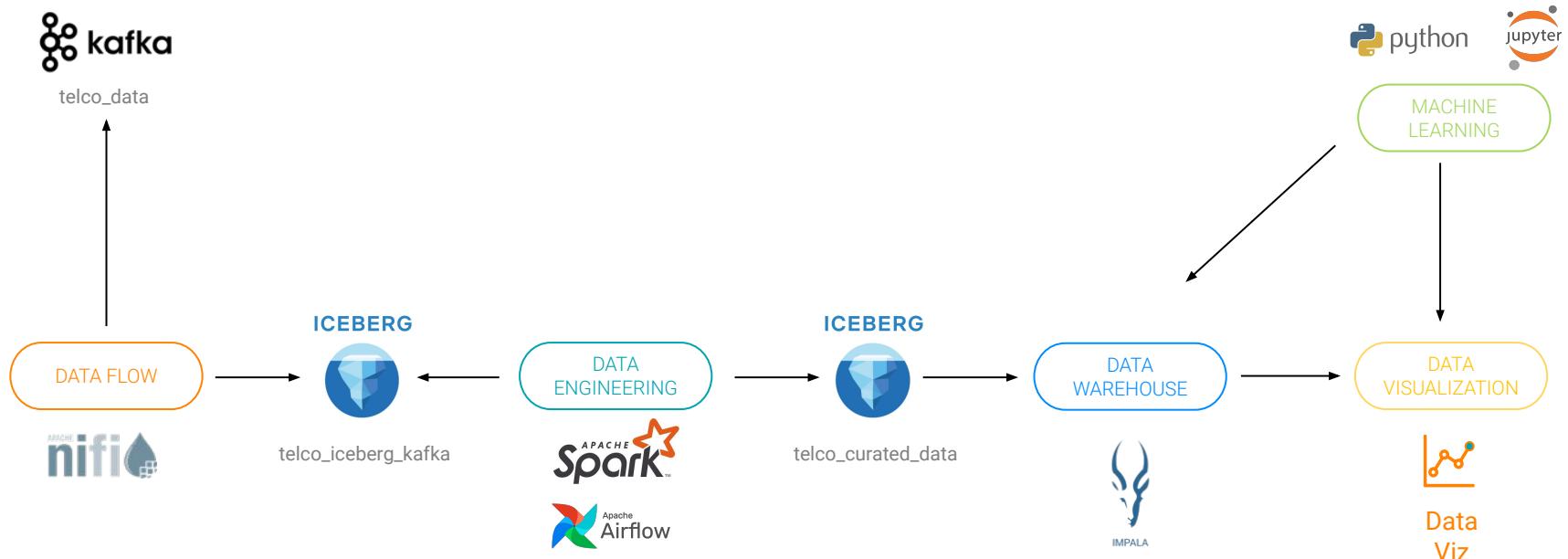


3000 Feet Use Case Description

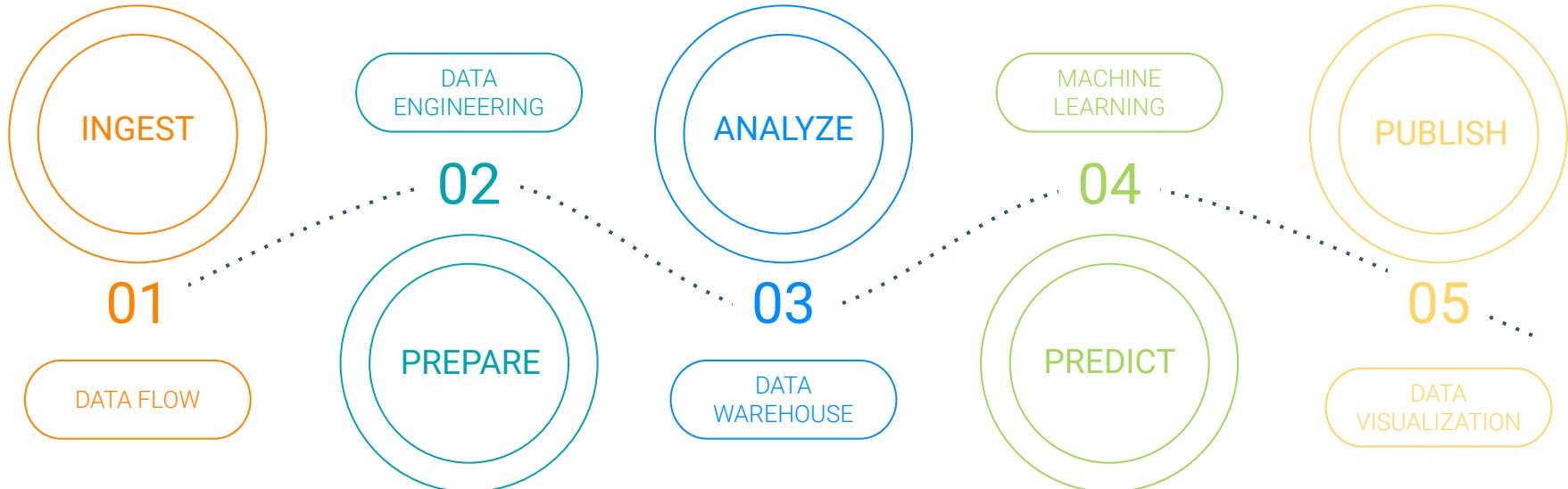
Customer Churn



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

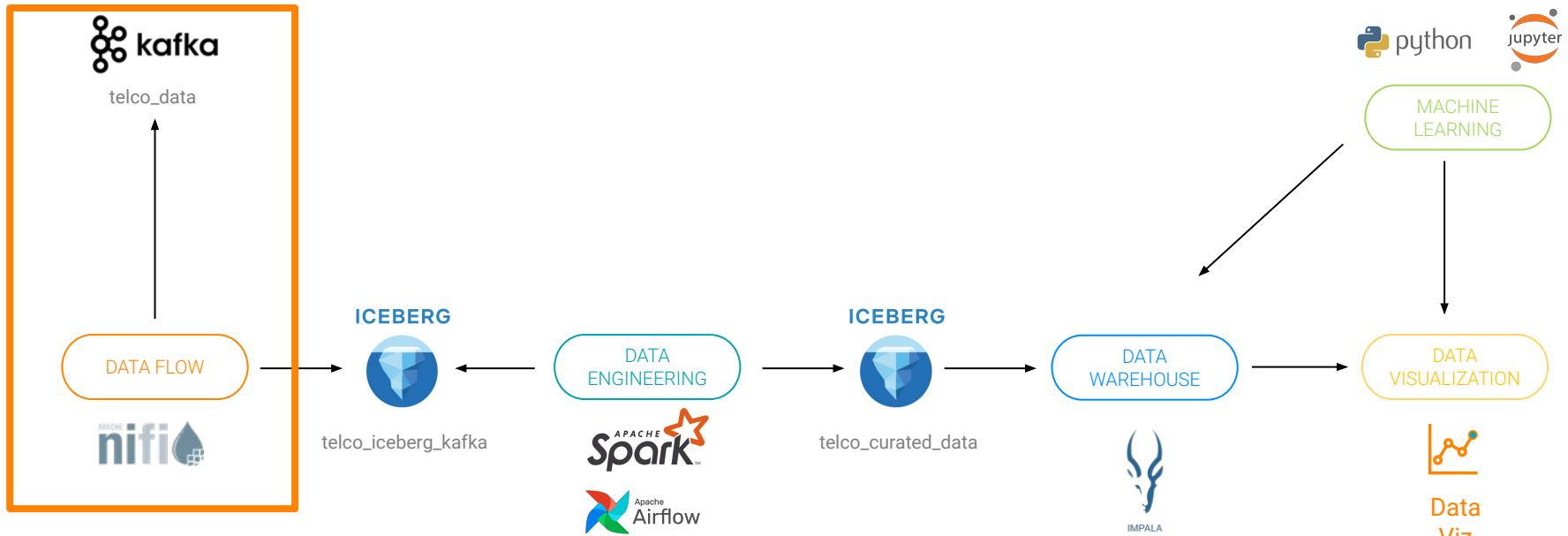
customerID	7590-VHVEG	5575-GNVDE	3668-QPYBK	7795-CFOCW	9237-HQITU	9305-CDSKC
gender *	F	M	M	M	F	F
SeniorCitizen	0	0	0	0	0	0
Partner *	Y	N	N	N	N	N
Dependents *	0	0	1	0	0	0
tenure	1	34	2	45	2	8
PhoneService	No	Yes	Yes	No	Yes	Yes
MultipleLines	No phone service	No	No	No phone service	No	Yes
InternetService	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

* Attributes to be enriched

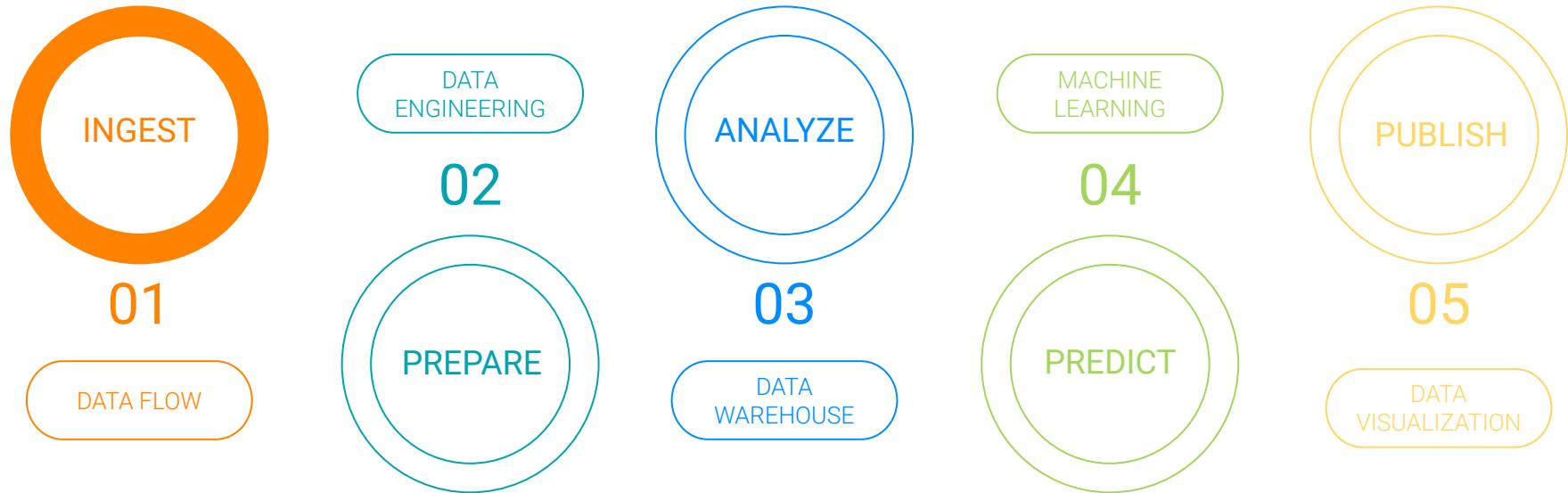
LAB 1:

Data Flow

ANALYTICS ACROSS THE DATA LIFECYCLE



ANALYTICS ACROSS THE DATA LIFECYCLE



CLOUDERA
SDX

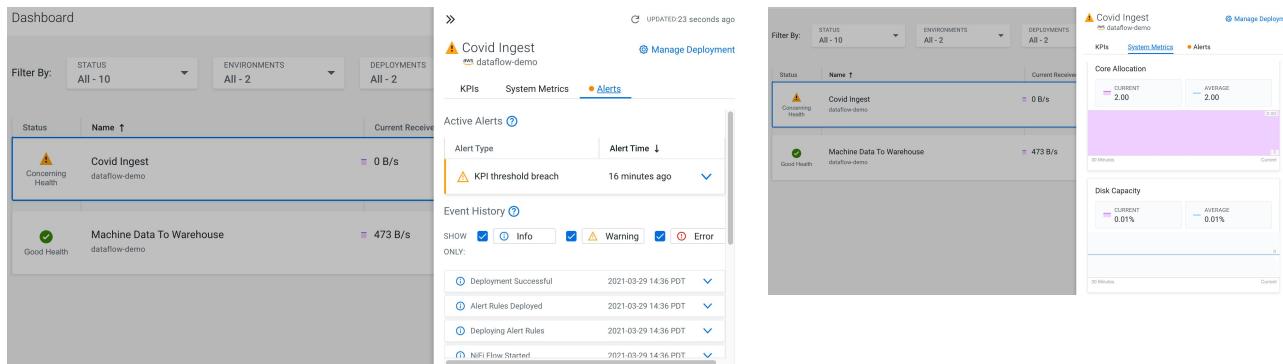
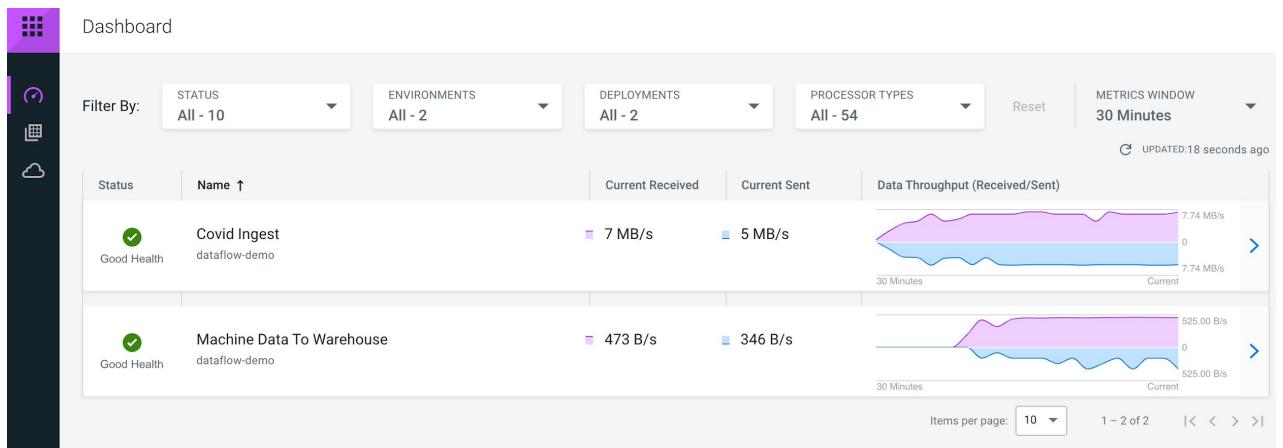
SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

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

DATA FLOW

- Easy Flow deployment
- Auto-scaling, resource isolation and cost-optimization
- Flow Catalog, central repository for flow definitions
- ReadyFlow Gallery, cover most common data flow use cases
- Easy maintenance/upgrades
- Advance monitoring capabilities
- Multiple data sources and sink operations



LAB 1 – COMPONENTS

A streaming message platform. It is designed to be high performance, highly available, and redundant, ideal for real-time and streaming applications



LAB 1 – COMPONENTS

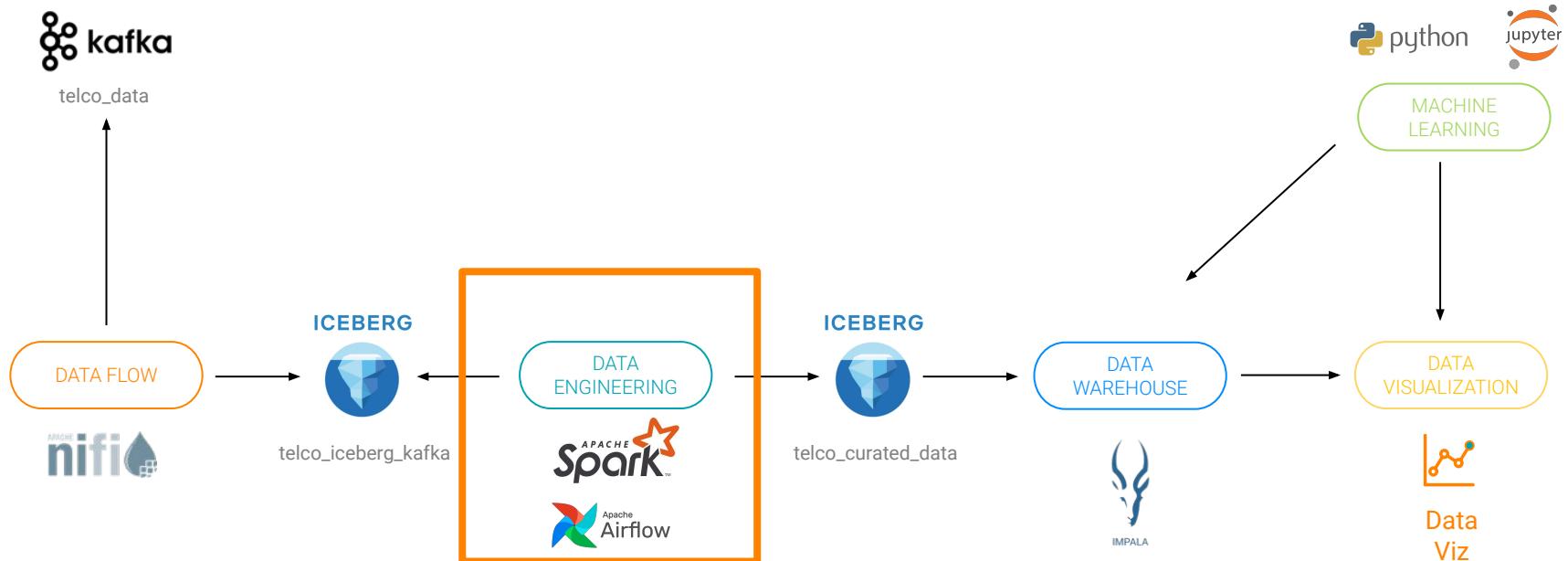


Apache NiFi is an open source software for automating and managing the flow of data between systems. It is a powerful and reliable system to process and distribute data. It provides a web-based User Interface for creating, monitoring, & controlling data flows.

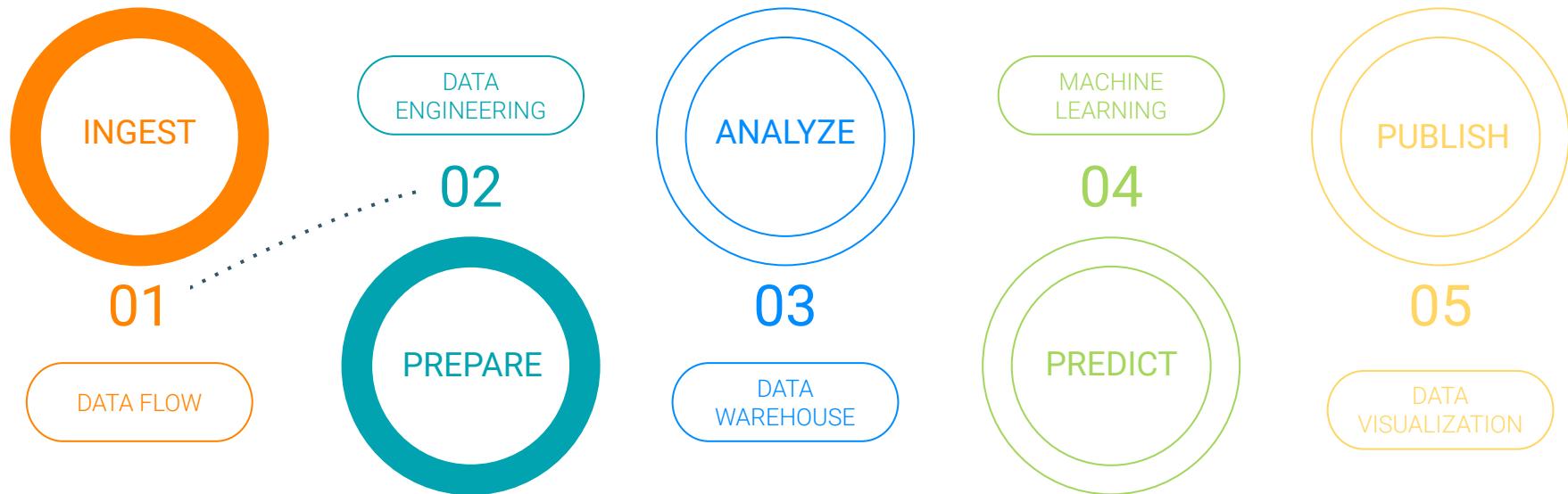
Some of the features includes **data provenance**, **extensible**, **secure** and others.

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

DATA ENGINEERING

- Streamlined service for scheduling, monitoring, debugging, and promoting data pipelines quickly & securely
- Inherited governance
- Deliver data pipelines to other CDP services (CDW, CML, etc)
- Portable and flexible
- Complete Data pipeline management

The screenshot displays the Cloudera Data Engineering interface. On the left, under 'Environments [3]', there are three entries: 'pse-721-cdp-env' (Enabled, 1 node, 8.0 CPU, 31 GB Memory), 'demo-aws-1' (Enabled, 1 node, 16 CPU, 61 GB Memory), and 'pse-aws-demo-cdp-env' (Enabled, 0 nodes, 0 CPU, 0 MB Memory). Below these is a button 'Enable new CDE'. On the right, under 'Virtual Clusters / demo-aws-1 [3]', there are three running workloads: 'Fraud-ETL-Dev-Workload' (7 pods, 3.7 CPU, 7 GB Memory, 0 jobs), 'HeavyETL-Workload' (7 pods, 3.7 CPU, 7 GB Memory, 0 jobs, with a line chart showing CPU usage spikes), and 'SalesOps-Analytics-Workload' (7 pods, 3.7 CPU, 7 GB Memory, 0 jobs).

LAB 2 – COMPONENTS



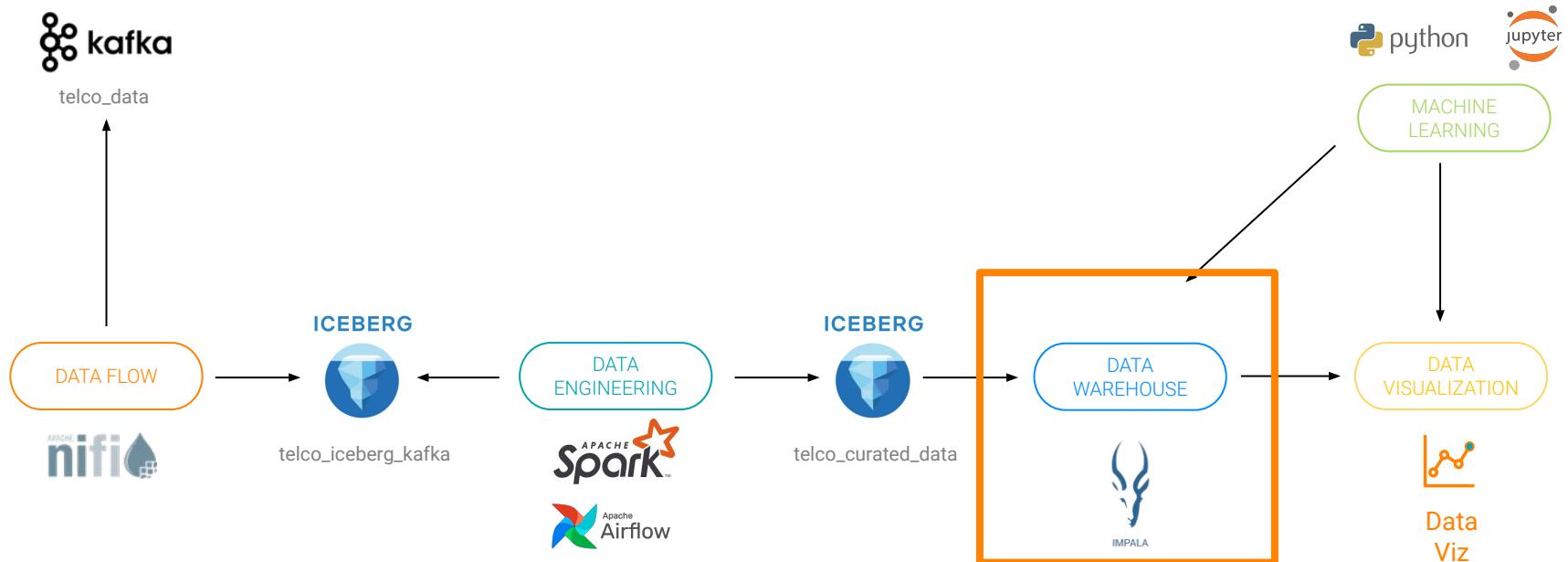
Apache Spark is an open-source, distributed processing system used for big data workloads. It utilizes in-memory caching, and optimized query execution for fast analytic queries against data of any size. It provides development APIs in Java, Scala, Python and R, and supports code reuse across multiple workloads—batch processing, interactive queries, real-time analytics, machine learning, and graph processing.

Open-source platform for developing, scheduling, and monitoring batch-oriented workflows. Airflow's extensible Python framework enables you to build workflows connecting with virtually any technology. A web interface helps manage the state of your workflows.

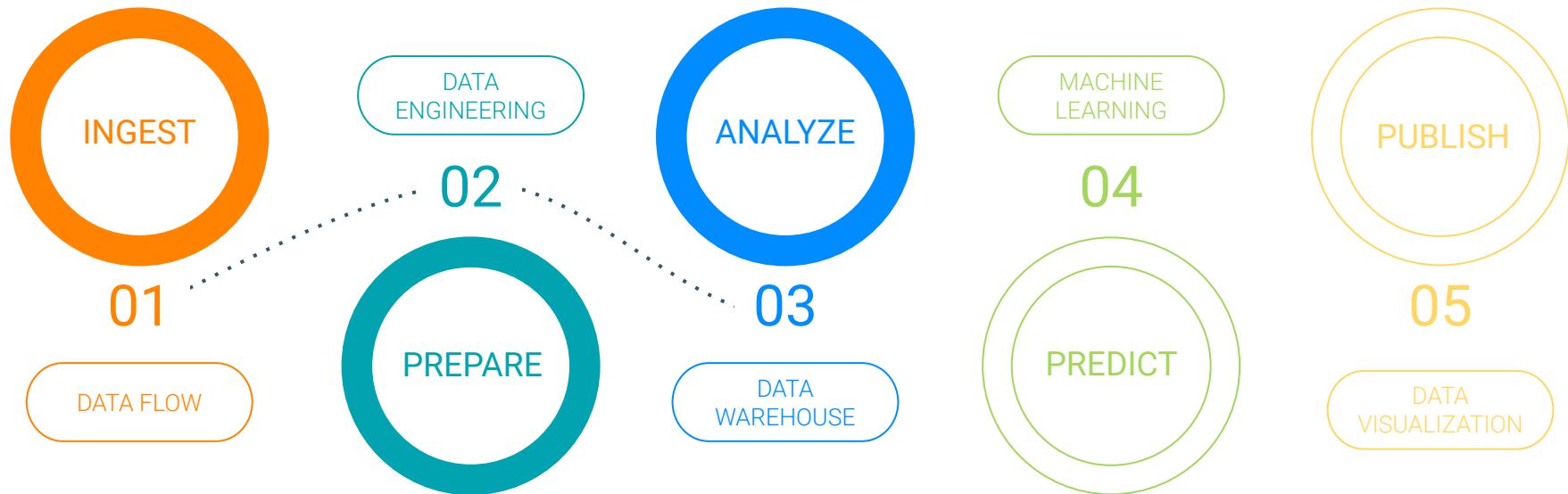


LAB 3: Data Warehouse

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

DATA WAREHOUSE

- Automated Capacity Planning
- Ease of Provisioning
- Auto-Scaling
- Resource Isolation
- High Concurrency
- Infrastructure optimized for Performance
- Choice of two DW engines: Hive and Impala, now Unified Analytics
- Fully integrated with Iceberg

The screenshot shows the Cloudera Data Warehouse X (DWX) interface. At the top right, it displays "DWX Version - 1.0.0.0-501". The main area is titled "Overview" and shows "Database Catalogs | 2". There are two entries: "it-demo-3-new" (Running) and "it-demo-3-default" (Running). Both entries have 2 databases, 16 GB memory, and 1 virtual warehouse. To the right, there is a section titled "Virtual Warehouses | 3" listing "prasanth" (Stopped), "it-warehouse-new" (Stopped), and "it-warehouse" (Stopped). Each entry has 0 nodes, 10 total cores, and 40 GB total memory, with a "HIVE" type indicated.

Virtual Warehouse	Status	Type	Nodes	Total Cores	Total Memory
prasanth	Stopped	HIVE	0	10	40 GB
it-warehouse-new	Stopped	HIVE	0	10	40 GB
it-warehouse	Stopped	HIVE	0	10	40 GB

LAB 3 – COMPONENTS



Hue is the open-source analytics workbench designed for fast data discovery, intelligent query assistance, and seamless collaboration. Bridge the gap between IT and the business for trusted self-service analytics.

Impala provides high-performance, low-latency SQL queries on data storage layer. The fast response for queries enables interactive exploration and fine-tuning of analytic queries.



IMPALA

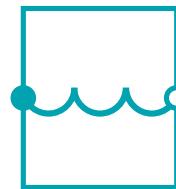
LAB 3 – COMPONENTS



DATA WAREHOUSE



DATA LAKE



LAKEHOUSE

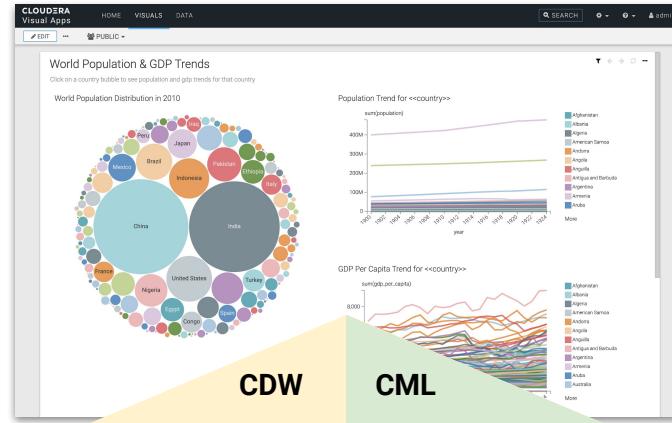


Highly curated
Precise
Complete
Well governed

Varying curation
Degrees of accuracy
Semi, unstructured
High granularity

Explore and discover
Trends and statistics
Predict and learn
Rapid time to insight

LAB 3 – COMPONENTS



CDW



CML



Real Time Event Store



Real Time Data Mart



Data Discovery & Exploration



Data Stream Analytics
(Coming soon)



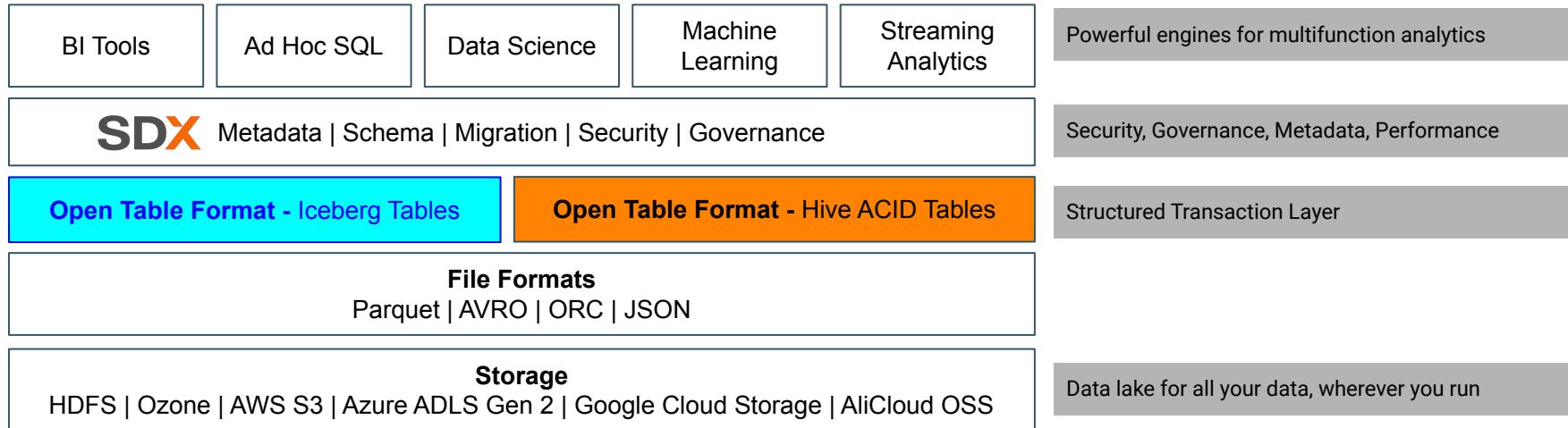
Predictive Models



Data Science

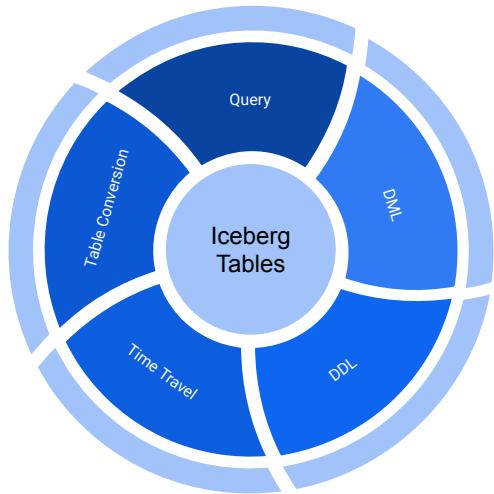


LAB 3 – COMPONENTS



* JSON is not supported by iceberg

LAB 3 – COMPONENTS



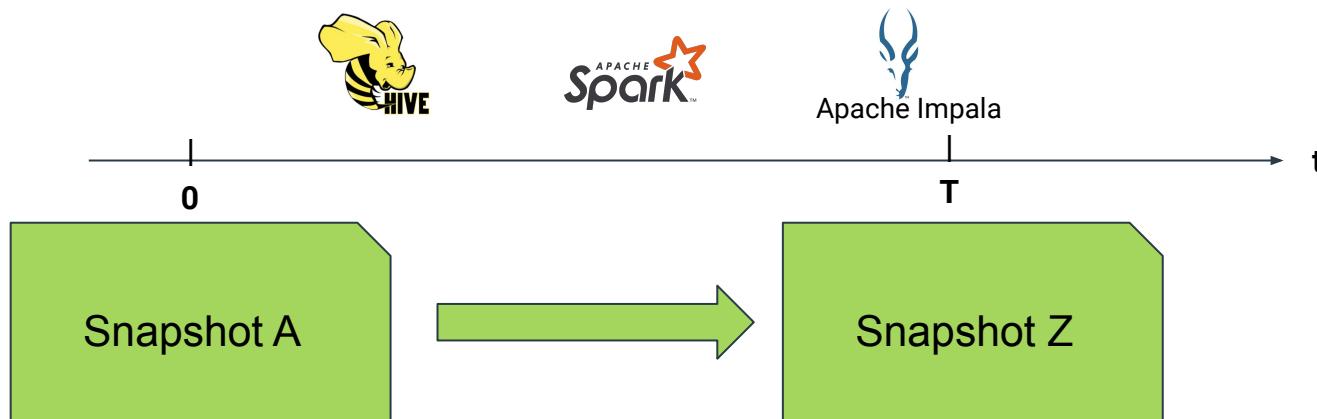
Rich set of SQL commands are developed for **Hive**, **Impala** and **Spark** to

- Create and manipulate database objects
- Run Queries
- Load data into tables
- Modify data in tables
- Perform Time Travel operations
- Convert to Iceberg tables

LAB 3 – COMPONENTS



Time Travel



Standard SQL operations:

- Queries
 - DDL
 - DML

Time Travel operations:

- SELECT ... AS OF ...

LAB 3 – COMPONENTS



Easy
conversion/adoption

1. Hive table migration:

```
ALTER TABLE tbl SET_TBLPROPERTIES  
('storage_handler'='org.apache.iceberg.mr.hive.Hiv  
eIcebergStorageHandler')
```

2. Spark 3:

a. Import Hive tables into Iceberg

```
spark.sql("CALL  
<catalog>.system.snapshot('<src>', '<dest>')")
```

b. Migrate Hive tables to Iceberg tables

```
spark.sql("CALL  
<catalog>.system.migrate('<src>')")
```



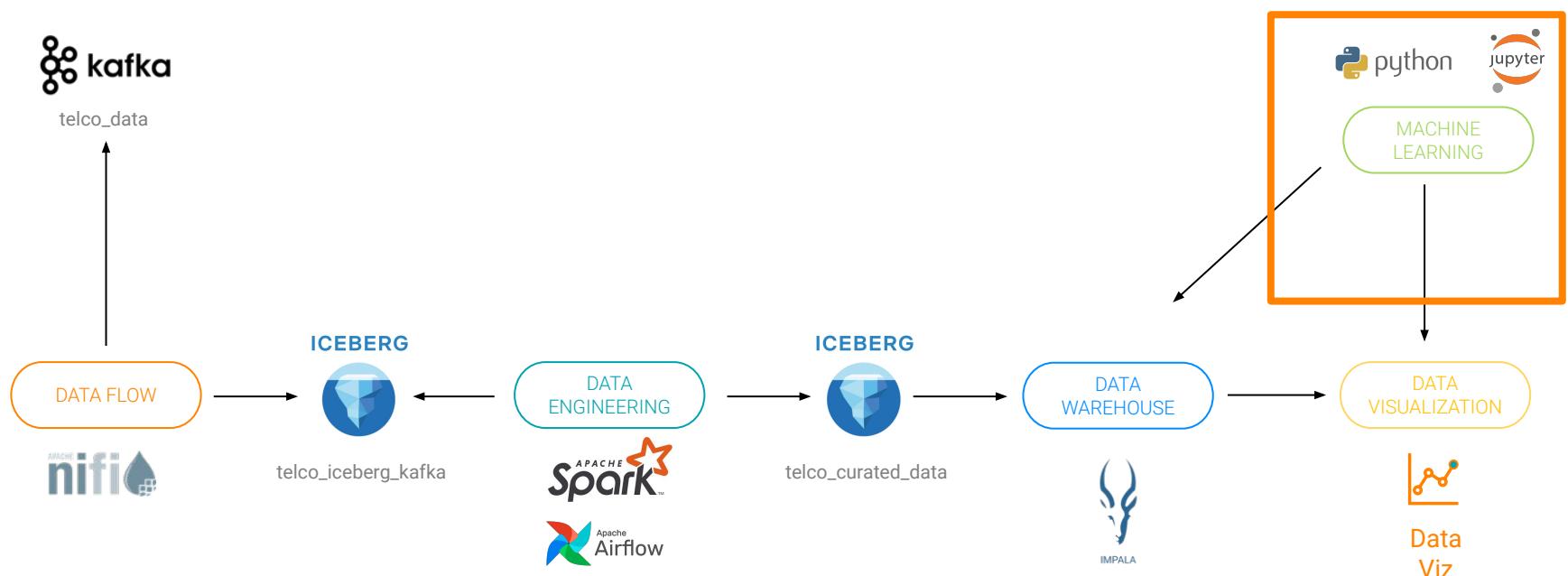
SDX

Iceberg Tables

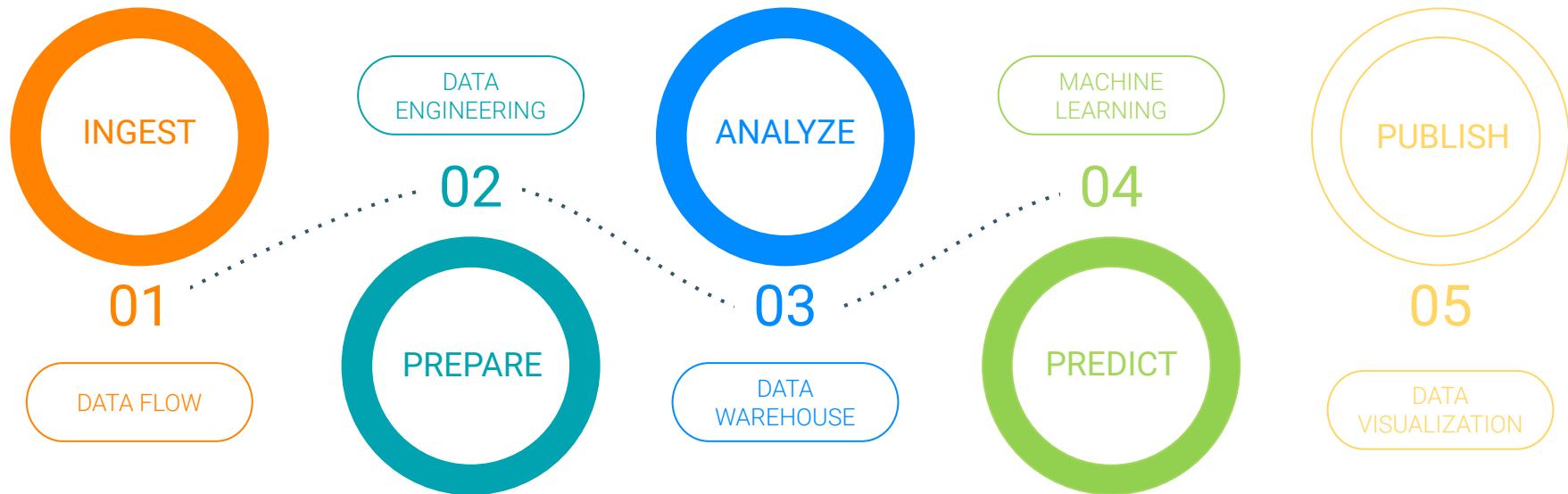
LAB 4:

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

MACHINE LEARNING

- ML Workspaces for teams without waiting
- Self-service governed data access
- Data scientists' preferred, open tools
- Elastic, auto-suspending resources
- Comprehensive, cohesive UX for end-to-end ML including DE
- Portable and consistent

The screenshot shows the Cloudera Machine Learning interface. On the left is a dark sidebar with a navigation menu:

- Projects
- Sessions
- Experiments
- Jobs
- Models
- Settings

The main area is titled "Projects" and displays various metrics:

Projects	Sessions	Jobs	Models	Experiments	Users
5	11	34	31	12	15

Below the metrics are three line charts showing resource usage over time:

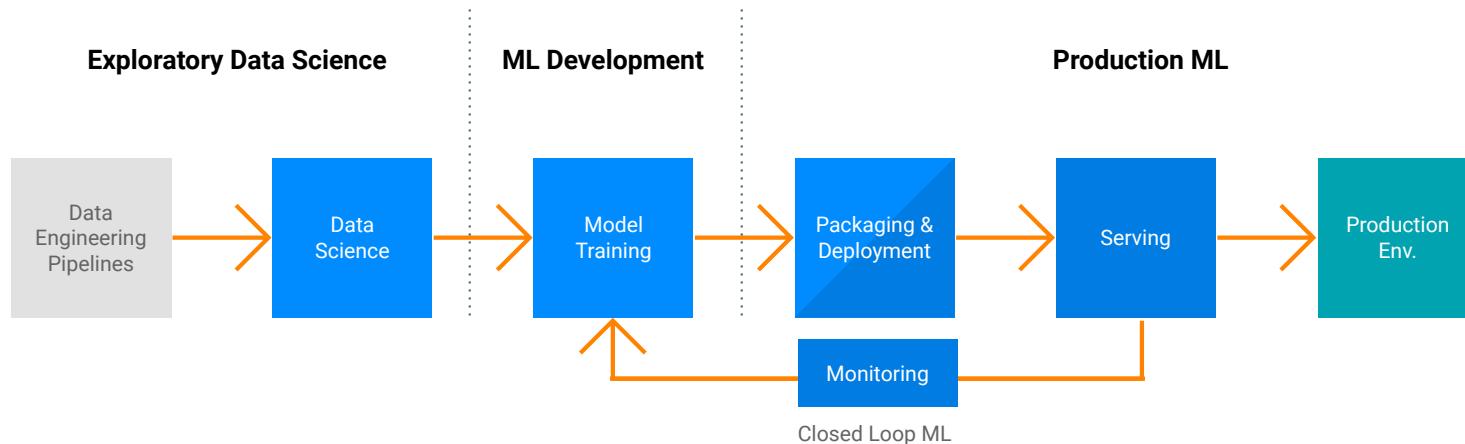
- vCPU: 24/36
- GPU: 12/24
- Mem: 24

At the bottom is a table titled "Projects" with the following data:

Projects	Created By	Last Updated	Active Sessions	Models
Admin 1	Test Admin	Jan 18, 2018, 10:54:31 AM	2	1
Payslip 2015	Danny	Jan 16, 2018, 04:54:31 PM	3	12
Sales Model Training	Jon	Sep 10, 2017, 03:31:00 PM	0	3
Team Sales 2019	Tynion	May 16, 2017, 09:45:56 AM	1	10
Team Marketing 2019	Robert	May 06, 2017, 10:31:22 AM	5	5

CML and Model Lifecycle

CML follows the 3 phases of the ML lifecycle



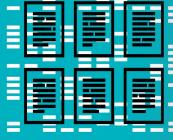
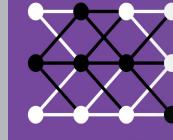
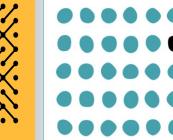
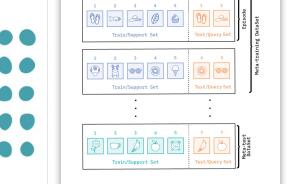
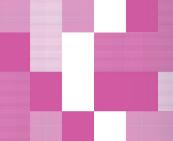
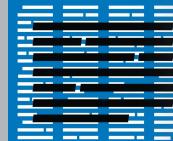
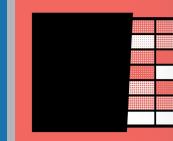
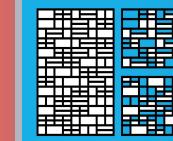
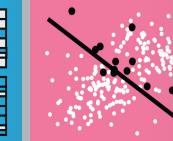
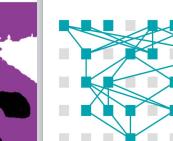
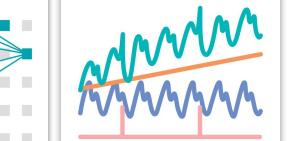
Guide & Flexibility

CLOUDERA
SDX

METADATA / SCHEMA / MIGRATION / SECURITY / GOVERNANCE

Choice & Governance

AMPs Seeded with Applied ML Research from Cloudera Fast Forward

Fast Forward Labs Natural Language Generation 	Fast Forward Labs Deep Learning: Image Analysis 	Fast Forward Labs Probabilistic Programming 	Cloudera Fast Forward Labs Semantic Recommendations 	Cloudera Fast Forward Labs Federated Learning 	Cloudera Fast Forward Labs Transfer Learning for Natural Language Processing 	CLOUDERA Fast Forward Deep Learning for Anomaly Detection 	CLOUDERA Fast Forward Meta-Learning 
Fast Forward Labs Probabilistic Methods for Realtime Streams 	Fast Forward Labs Summarization 	Fast Forward Labs Interpretability 	Cloudera Fast Forward Labs Multi-Task Learning 	Cloudera Fast Forward Labs Learning with Limited Labeled Data 	Cloudera Fast Forward Labs Deep Learning for Image Analysis 2019 Edition 	CLOUDERA Fast Forward Causality for Machine Learning 	CLOUDERA Fast Forward Structural Time Series 

[Preview all of our research here](#)

MLOps – Model Governance

Functional requirements



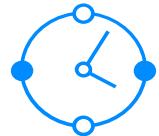
Repository

Registry of the models, versions, when it was deployed, data dependencies, how to run it...



Deploy

Get easily insights from the model in production, without re-writing the model



Monitorization

Getting the behaviour of the model according with defined business and statistical metrics with a certain periodicity



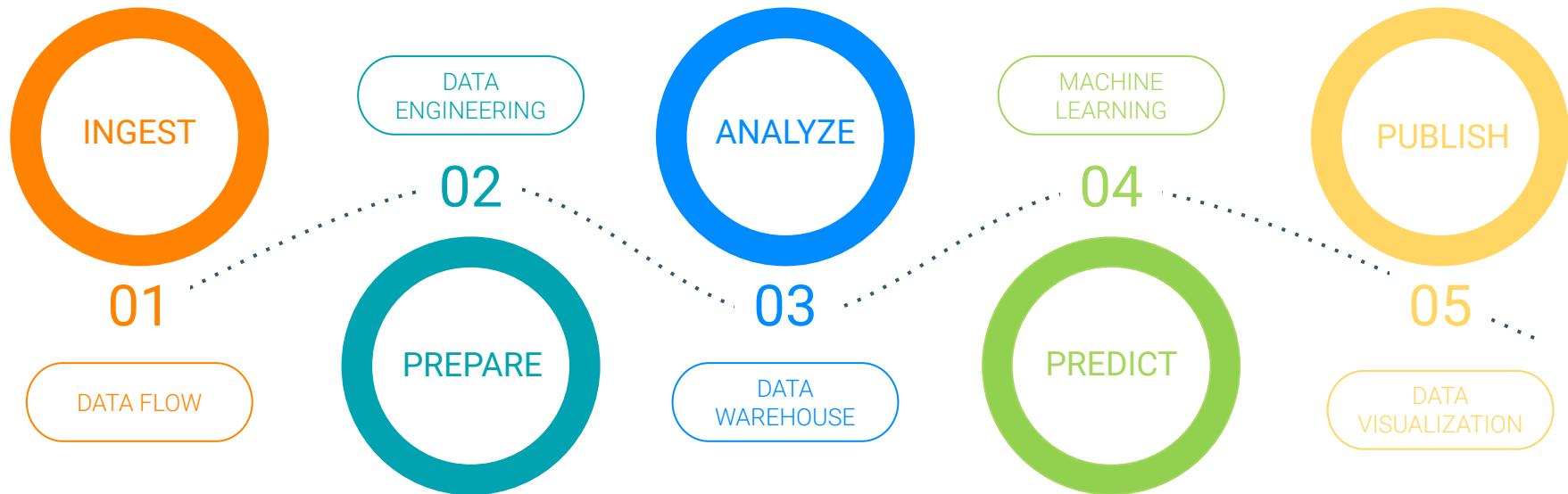
Automatization

The automatic accomplishment of the previous requirements would lead to a robust mlops methodology but with capability of define different strategies for the different projects

LAB 5:

Optional

ANALYTICS ACROSS THE DATA LIFECYCLE



CLOUDERA
SDX

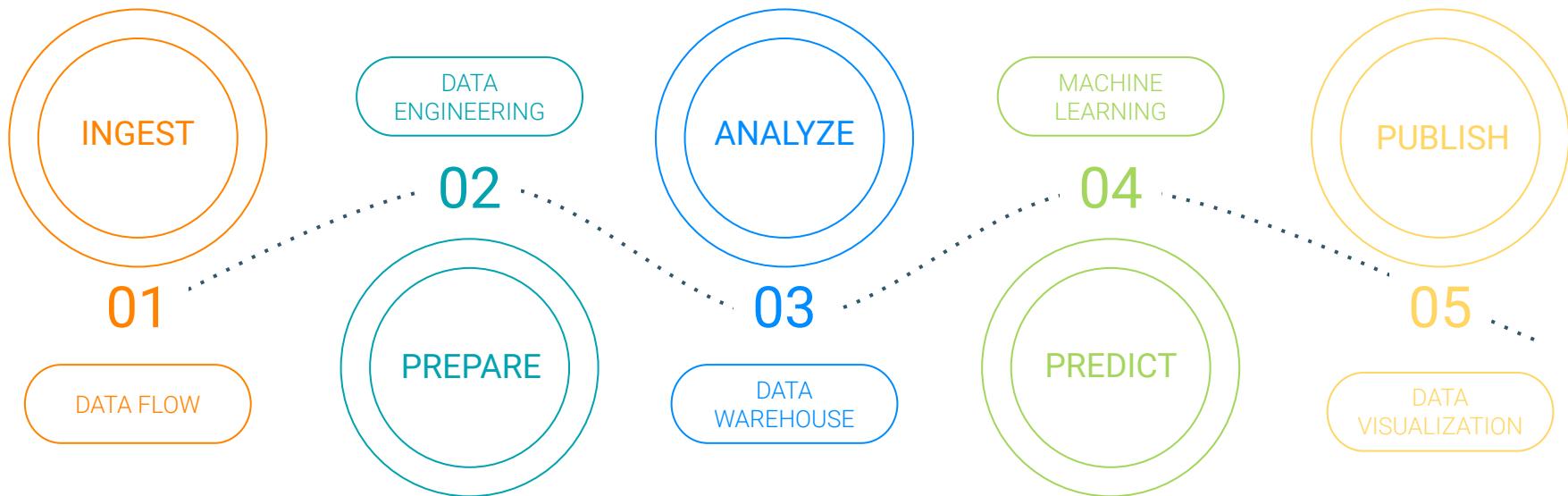
SECURITY | GOVERNANCE | LINEAGE | MANAGEMENT | AUTOMATION

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

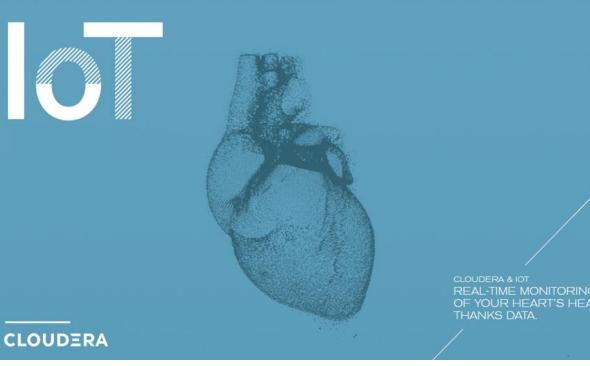
Next Step

- 18 octobre - Paris présentiel
 - Atelier de modernisation de workloads dans le cloud public
 - <https://rb.gy/k172s>
- 19 octobre - Webinaire virtuel
 - Comment traiter vos flux en temps réel?
 - <https://rb.gy/rwi66>
- 9 novembre - Webinaire virtuel
 - Open Data Lakehouse
 - <https://rb.gy/va04f>
- 16 Novembre - Paris présentiel
 - Atelier flux de données
 - <https://rb.gy/bj4ip>



ENTERPRISE DATA CLOUD

CLOUDERA



IoT

CLOUDERA

CLOUDERA Now

CLOUDERA



EDGE
2AI

CLOUDERA



CLOUDERA

CLOUDERA
HANDS-ON
EXPERIENCE

Thank you!