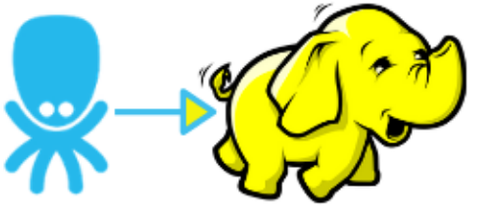
Big Data Platform @HCVN

(Nov-2020)



Deliverables

1. Big Data intro
2. Hosel Conf. 2020 takeaways
3. Big Data Platform (BDP) @HCVN
4. BIG Data demo

What is Big Data?

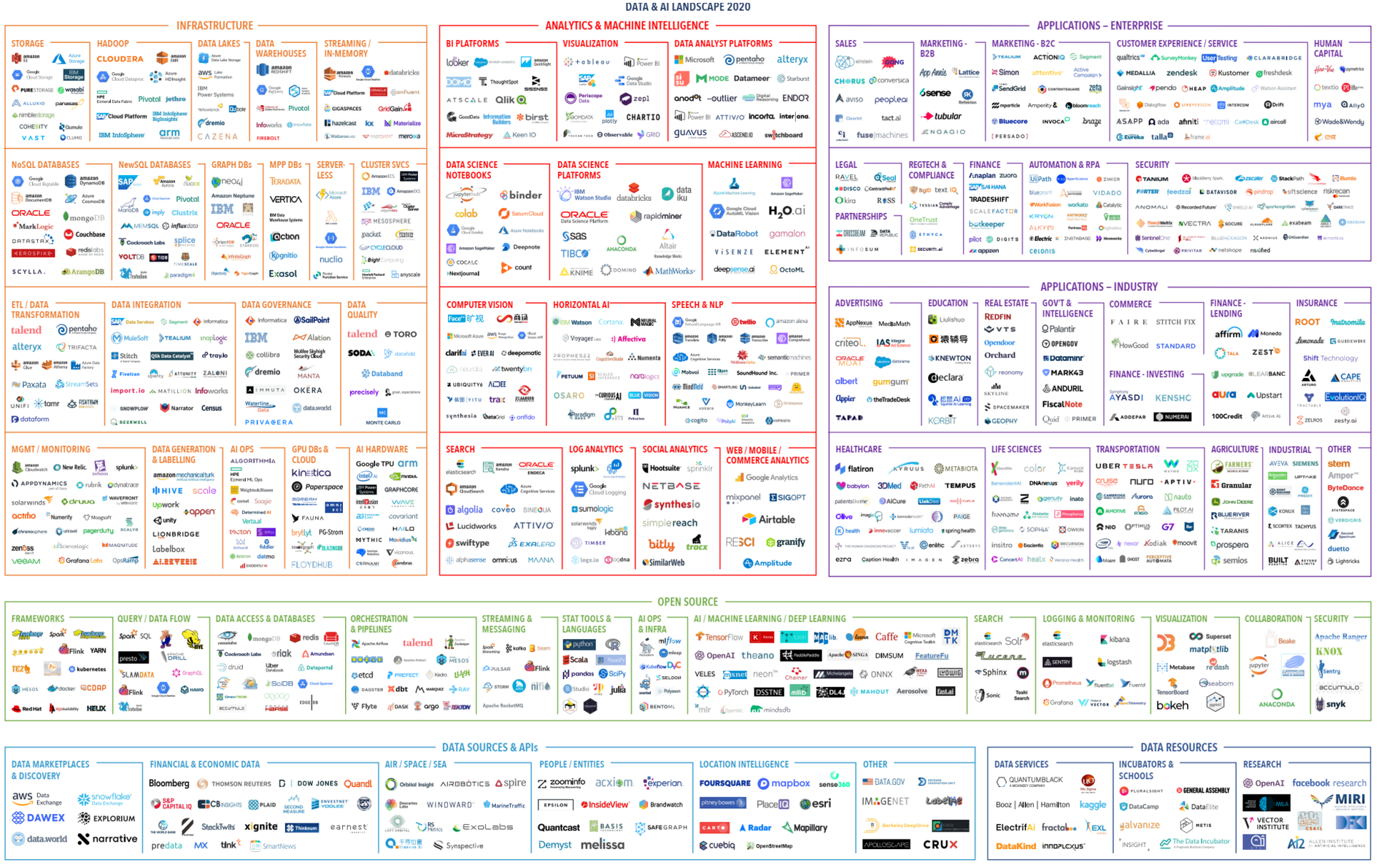
3Vs definition

- Volume
- Velocity
- Variety
- ...

Can’t definition

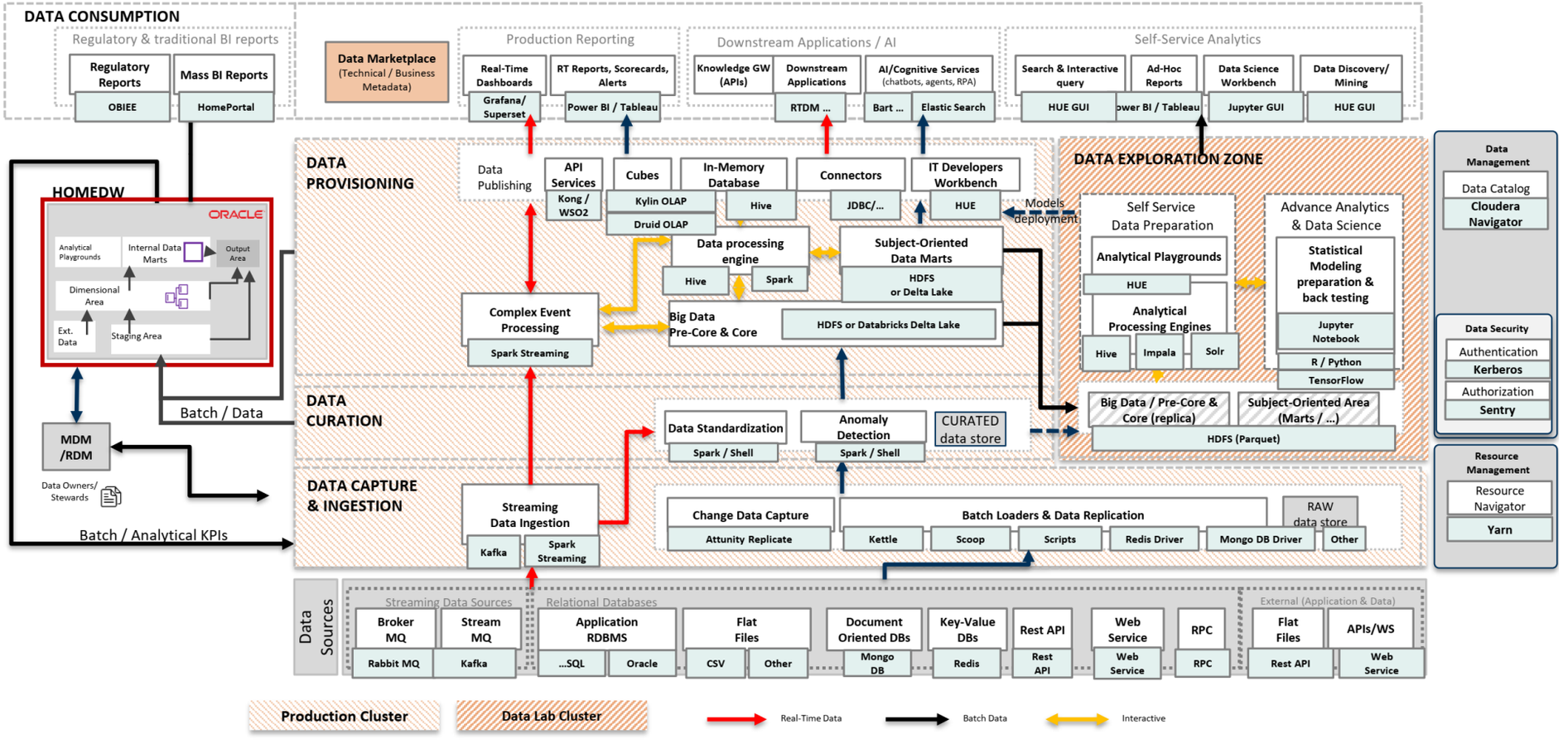
- The task brings down production database
- The task takes too long to run
- The task requires too many steps to complete
- …

Big Data landcape in 2020



Hosel Conf. 2020

Conceptual Data Architecture



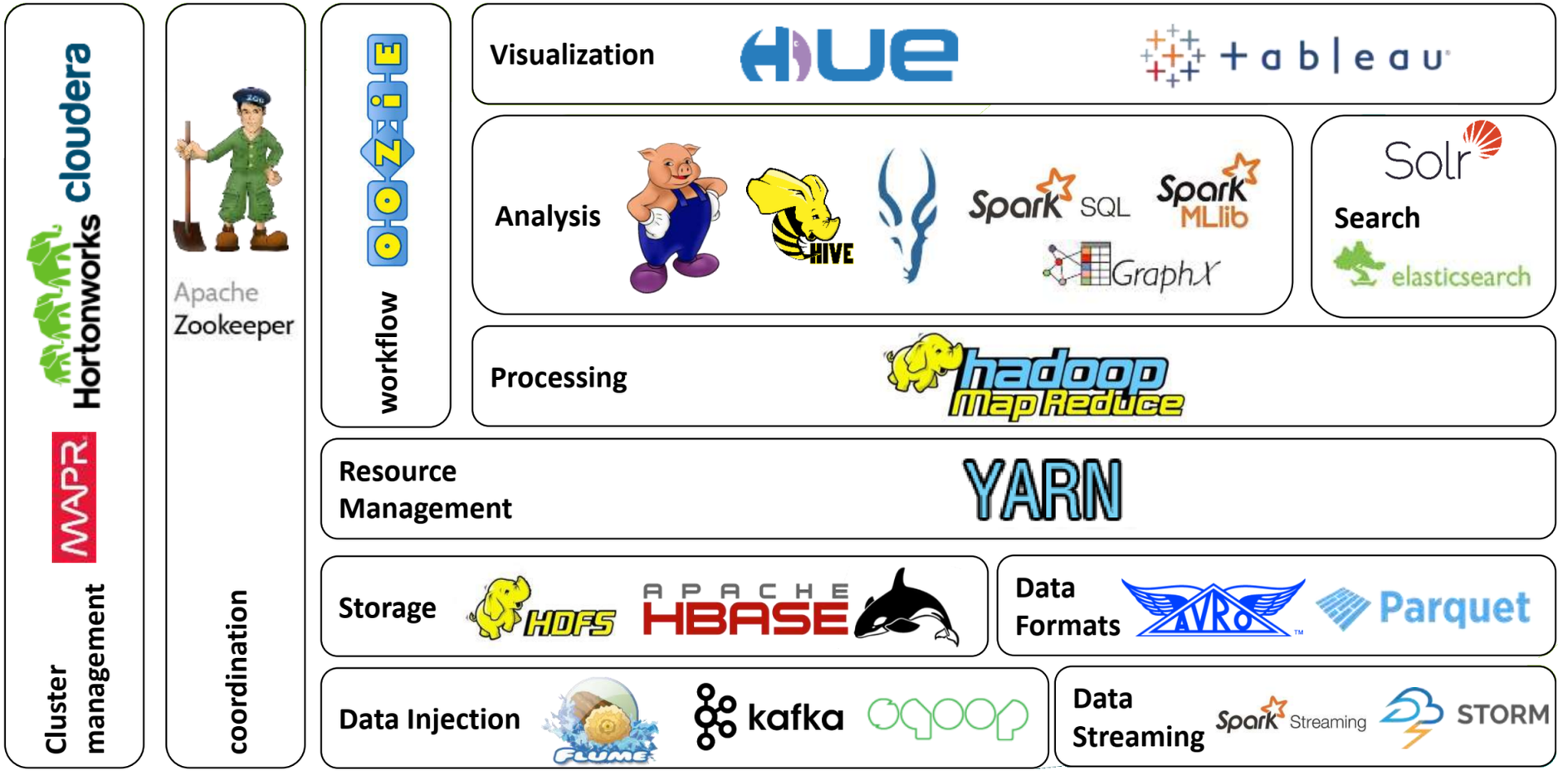
HoSel conference in Almaty, 4th to 6th February 2020

MS Stream >> HoSel 2020 Almaty Kazakhstan | Big Data

In-house solution

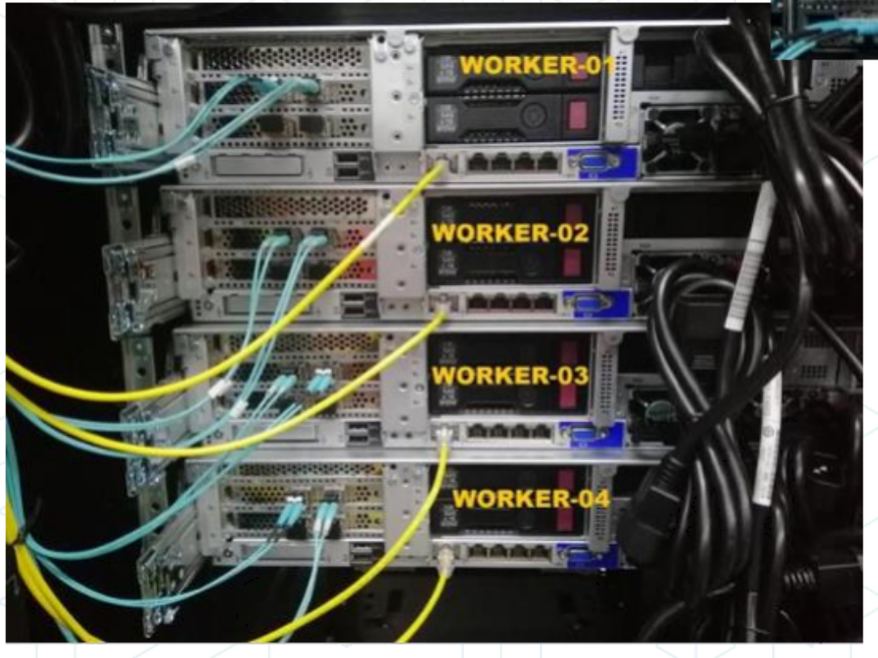
- Revise lessions learned from BDP in CN
- Stardardize group-wide Architect
- Guarantee better data governance
- Adopt open sources

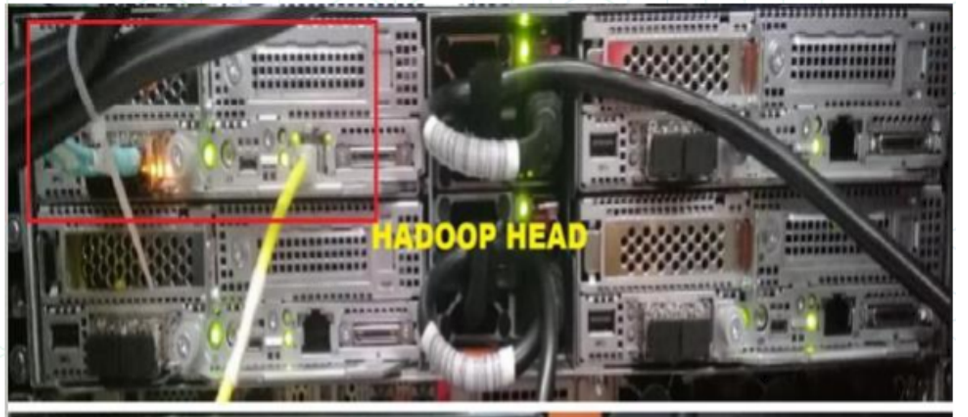
Hadoop Ecosystem



Hadoop baremetal cluster deloyed to Hosel countries

PH PROD cluster





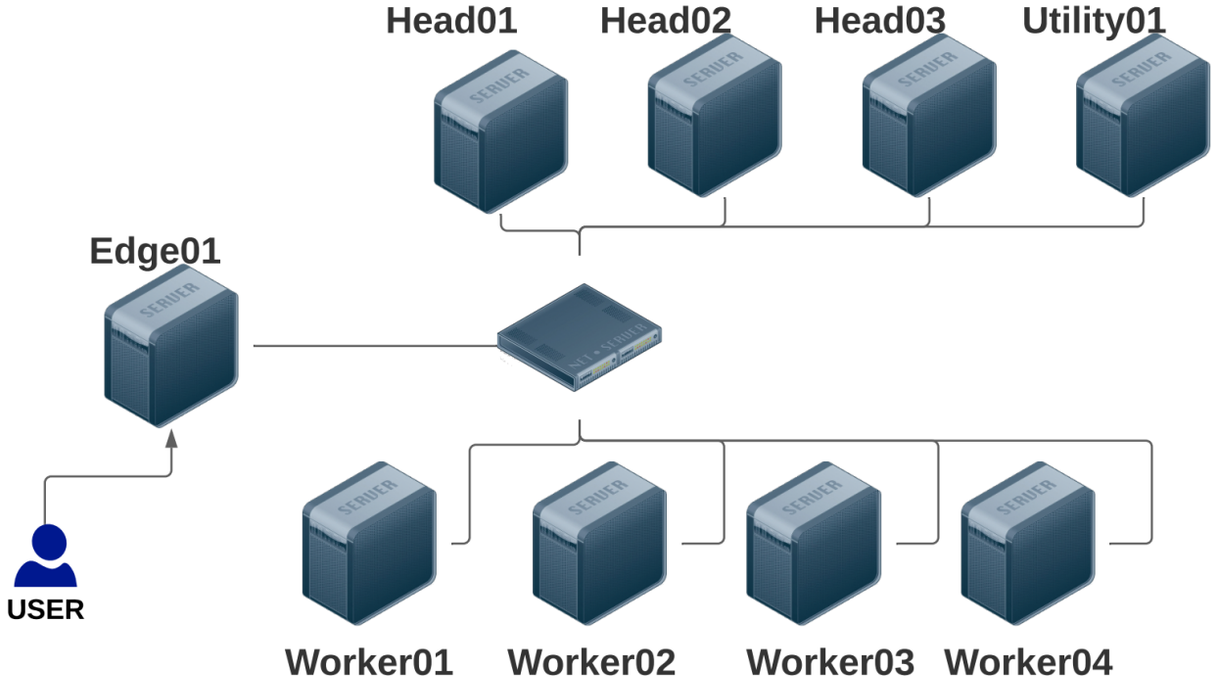
- Hosel countries: ID, IN, KZ, PH & VN
- VN is the last on in the list

BDP at HCVN

Keboola to Hadoop migration

- Work with EmbedIT on alternative components
- Speed up learning the zoo of Big Data
- 90% migration complete, target 100% by EOY
- Deliver BDP project requests from BU

Infrastructure



- **Edge**: user self-service access (Jupyter, NIFI, etc.)
- **Head**:
  - HDFS meta data
  - Other services (yarn, hue, etc.)
- **Worker**: store and process data.
- **Utility**: other services (Zeppelin, CMF, etc.)

💡 Easily scale horizontally by adding more worker nodes when more resources required

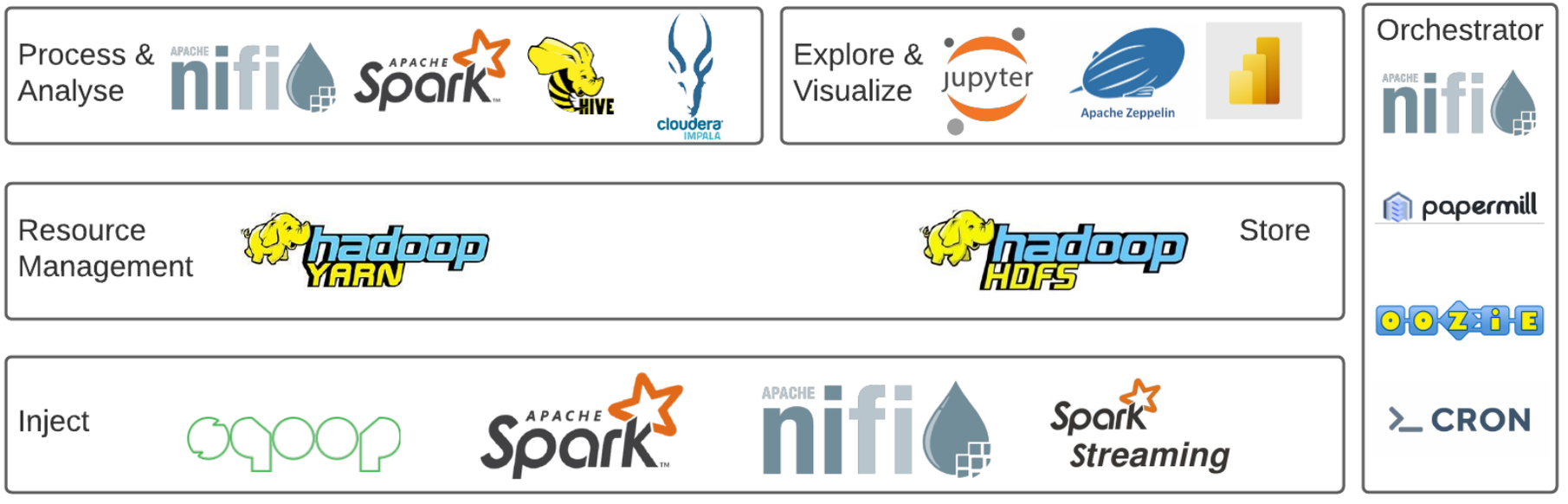
Infrastructure (cont)

Status	Name	Cores	Disk Usage	Physical Memory
✓	bdp-edge01-pdc.vn.prod	8	98 GiB / 429.7 GiB	46.2 GiB / 125.7 GiB
✓	bdp-head01-pdc.vn.prod	8	76.6 GiB / 149 GiB	57.1 GiB / 62.8 GiB
✓	bdp-head02-pdc.vn.prod	8	66 GiB / 149 GiB	26.9 GiB / 62.8 GiB
✓	bdp-head03-pdc.vn.prod	8	57.6 GiB / 149 GiB	23.8 GiB / 62.8 GiB
✓	bdp-utility01-pdc.vn.prod	8	81.7 GiB / 176.2 GiB	17.4 GiB / 31.3 GiB
✓	bdp-worker01-pdc.vn.prod	40	4.6 TiB / 28.8 TiB	38.6 GiB / 125.6 GiB
✓	bdp-worker02-pdc.vn.prod	40	5.3 TiB / 28.8 TiB	38.4 GiB / 125.6 GiB
✓	bdp-worker03-pdc.vn.prod	40	5.3 TiB / 28.8 TiB	40.3 GiB / 125.6 GiB
✓	bdp-worker04-pdc.vn.prod	8	5.2 TiB / 31.6 TiB	26.5 GiB / 62.8 GiB

- **Total CPUs**: 168
- **RAM**: 785GB
- **Storage**: 119TB

(Updated via Cloudera Manager on 10 Nov, 2020)

Focused stack



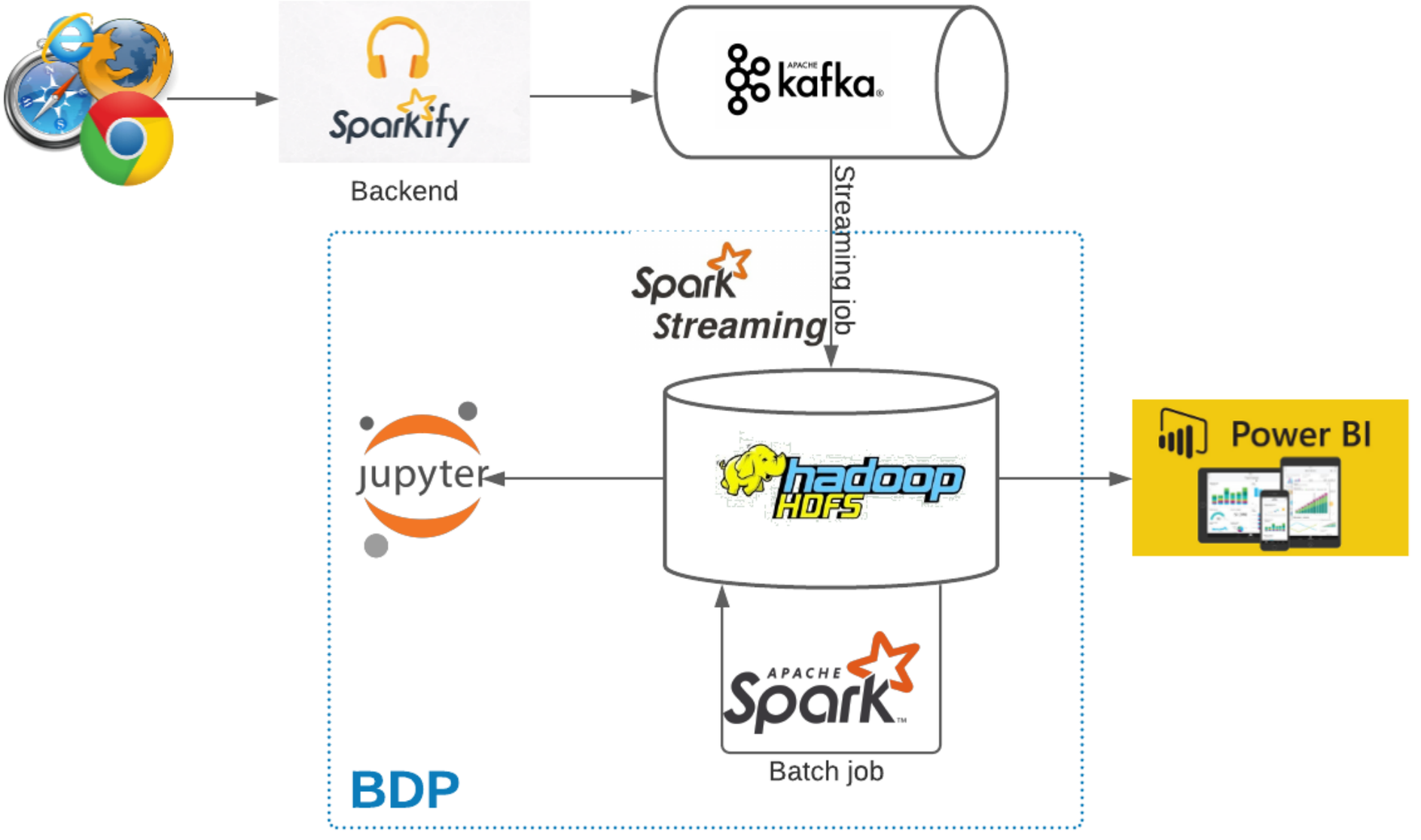
Quick access

- NIFI: Data flow management
- HUE: Hadoop assistant
- Spark/Streaming: Data process engine
- Jupyter/Zeppelin: web-based interactive development

Demo

Requirements

- [DE] Build a data pipeline utilize BDP components
- [DA] Build dashboard reporting user behavior metrics with PowerBI
- [DS] Build a model to predict churned users



Business questions

1. Which gender is more active?
2. Which level is more active (free or paid)?
3. Which factors (based on collected data) make users stop subcribing the service (churn)?

Summary

- When to use Big Data
- HC Group BDP strategy
- HCVN BDP overview
- Develop an end-to-end data project with BDP

Read more at

- HCI Big Data workspace
- ITBI Big Data workspace
- BDP Holsel Conference
- Demo source code

Thank you!