Managing Capacity @ LinkedIn



Anuprita Harkare

Site Reliability Engineer, LinkedIn


MY DOCTOR SAYS MY HAIR LOSS IS CAUSED BY ME WORRYING ABOUT DATA LOSS!

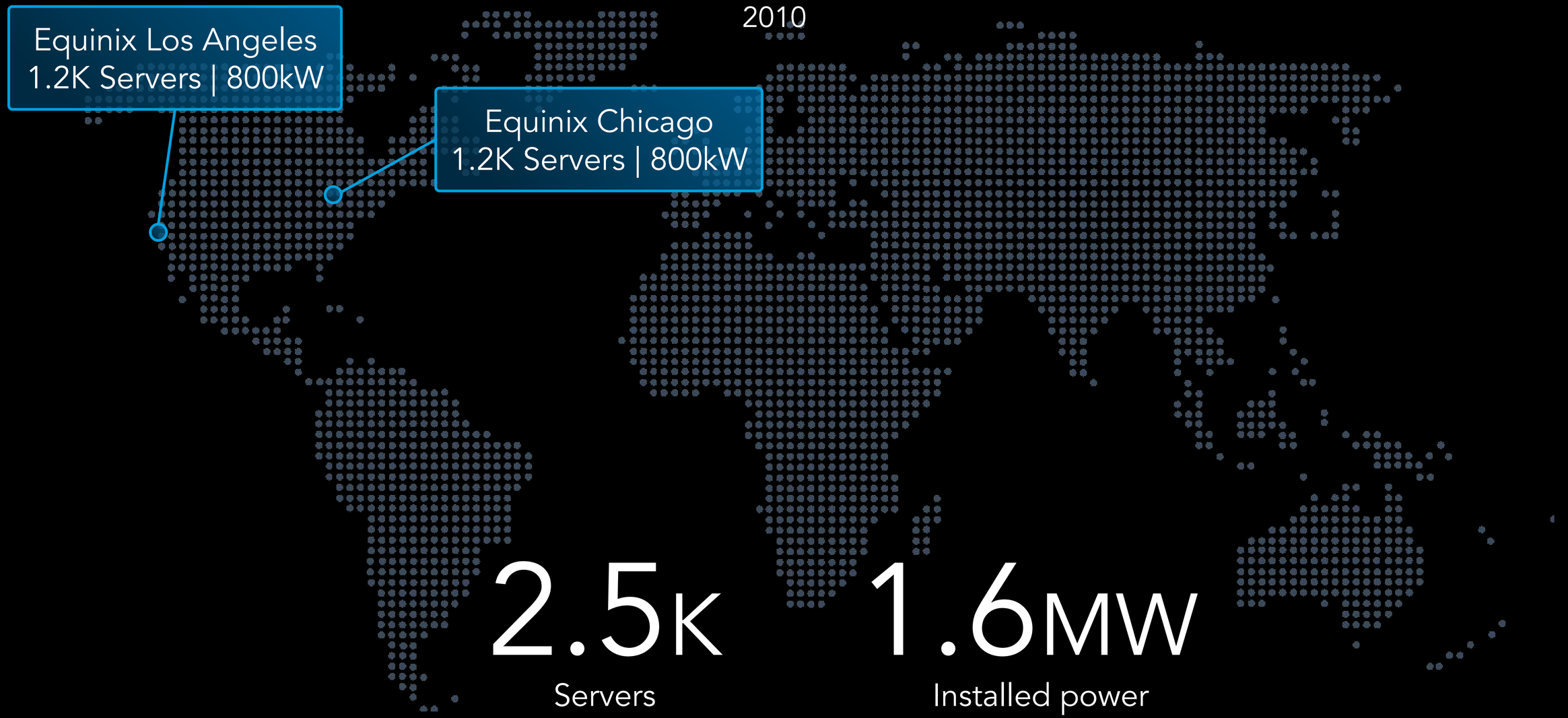


Agenda

- Linkedin’s data footprint
- The Data pipeline
- Underlying Components
- Tools & Metrics
- Q&A

Entire Production Footprint

2010

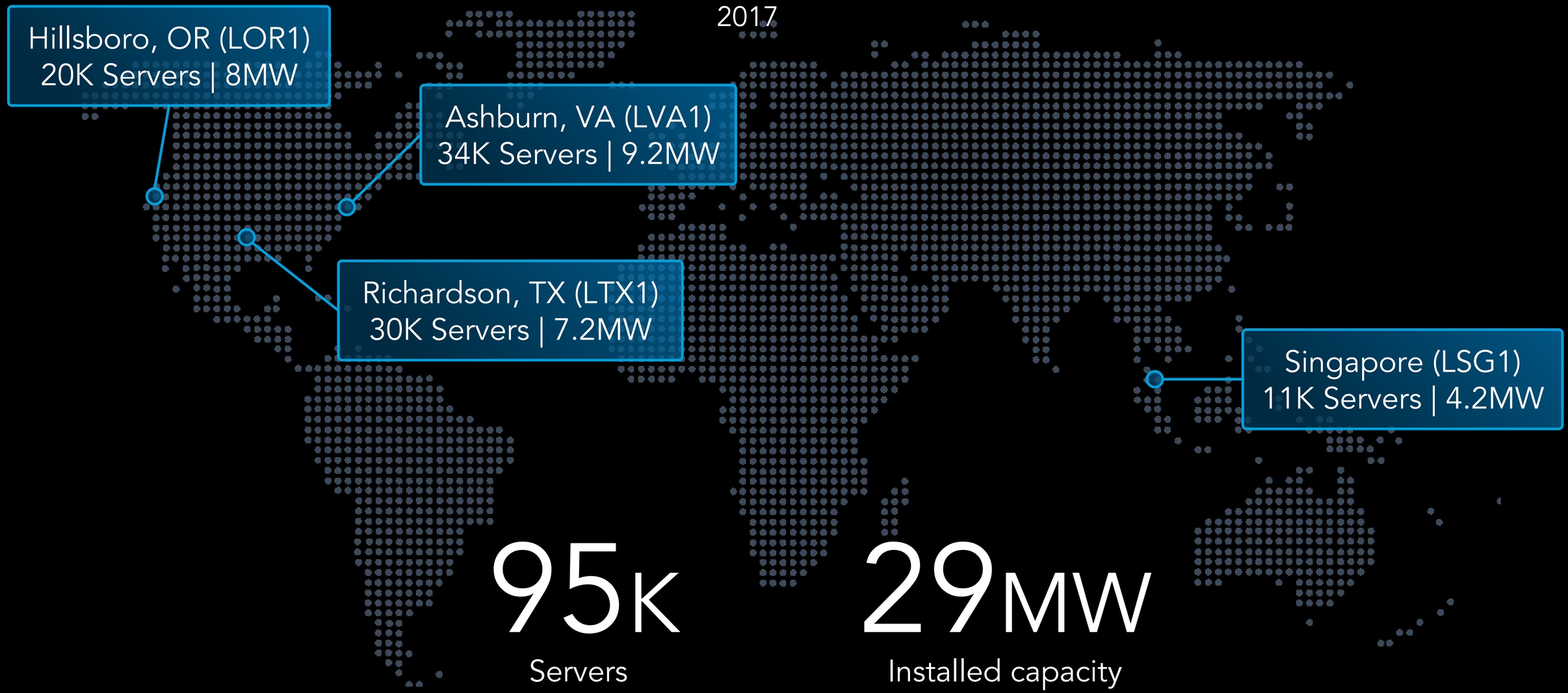


2.5K
Servers

1.6MW
Installed power

Production Application Footprint

2017



95K
Servers

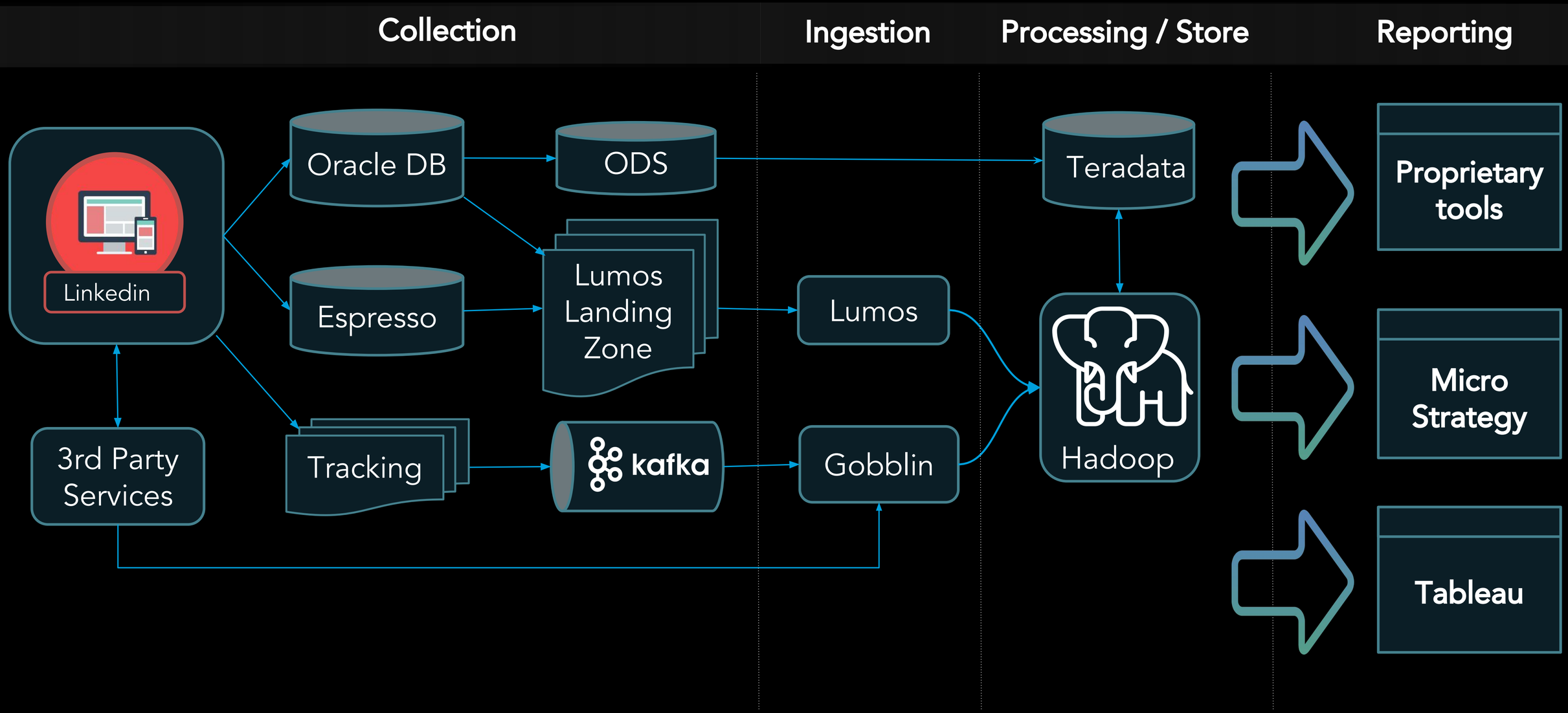
29MW
Installed capacity

Storage By the Numbers

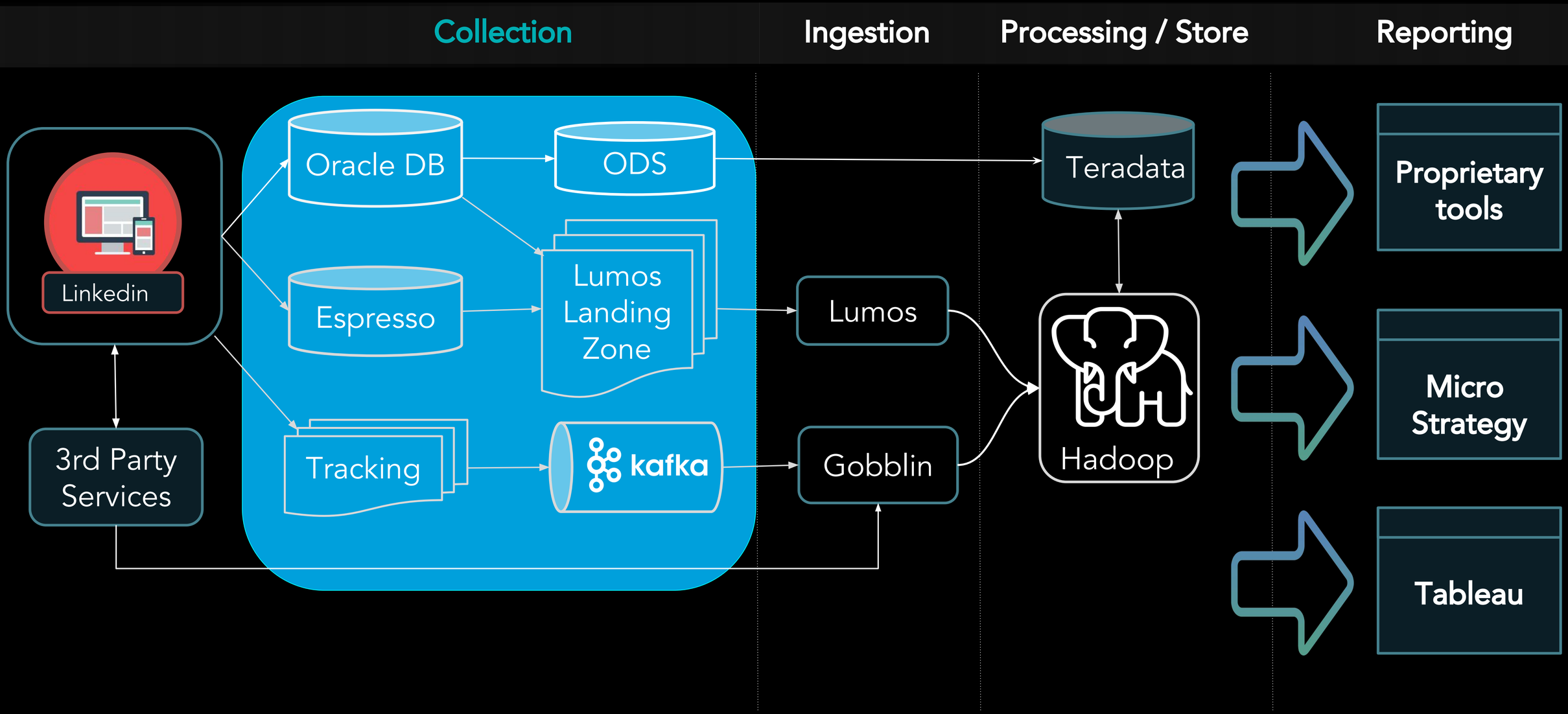
4.2PB	170PB	140TB	1.3PB
Espresso Storage	HDFS Storage	MySQL Storage	Oracle Storage

1600	2T
Kafka Brokers	Kafka messages per day

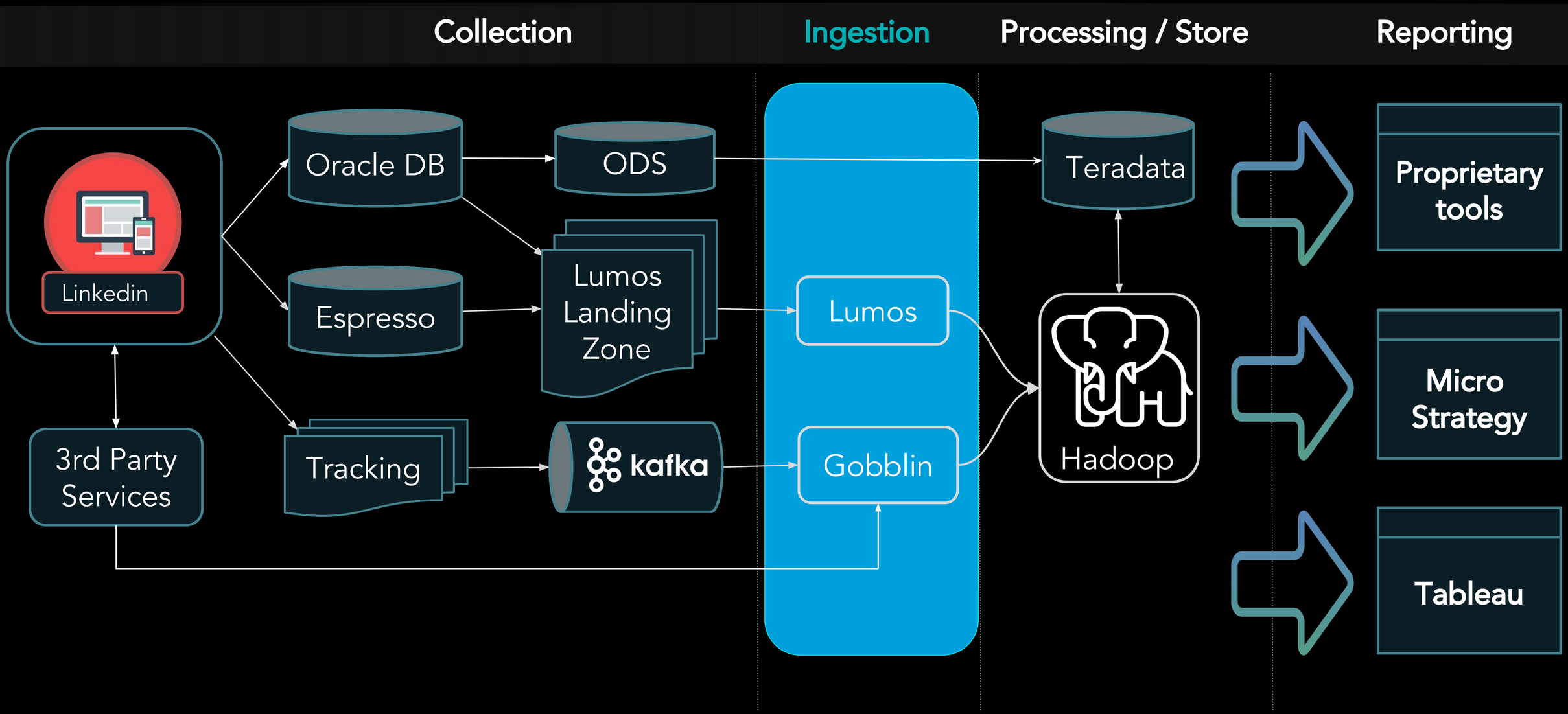
Linkedin's pipeline



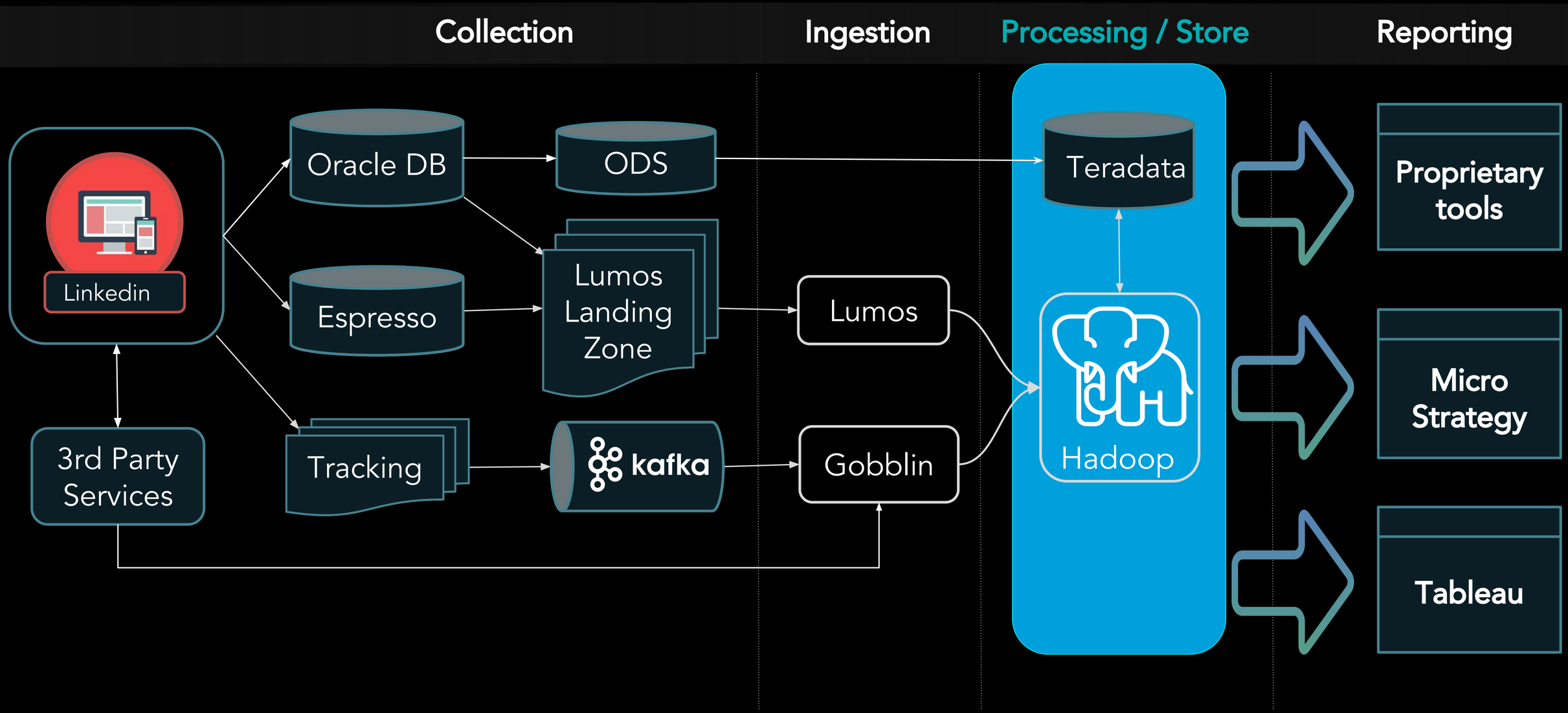
Linkedin's pipeline



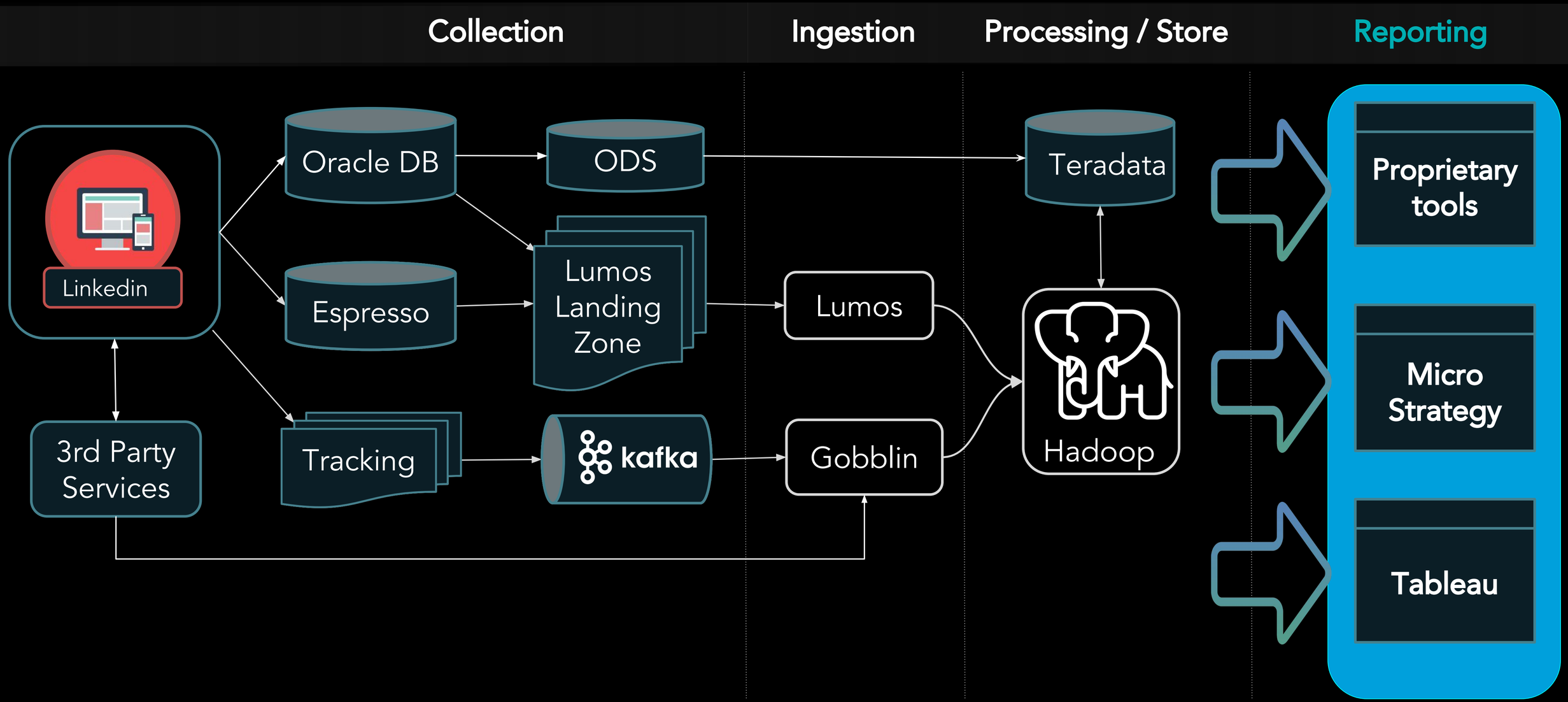
Linkedin's pipeline



Linkedin's pipeline



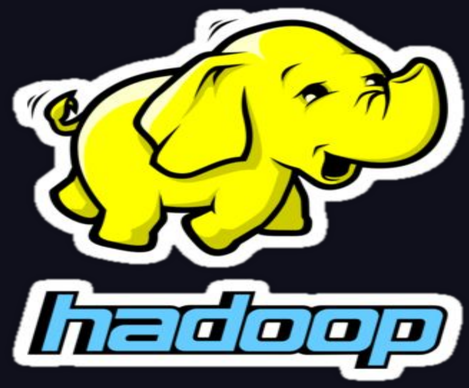
Linkedin's pipeline



Capacity planning for Major Components







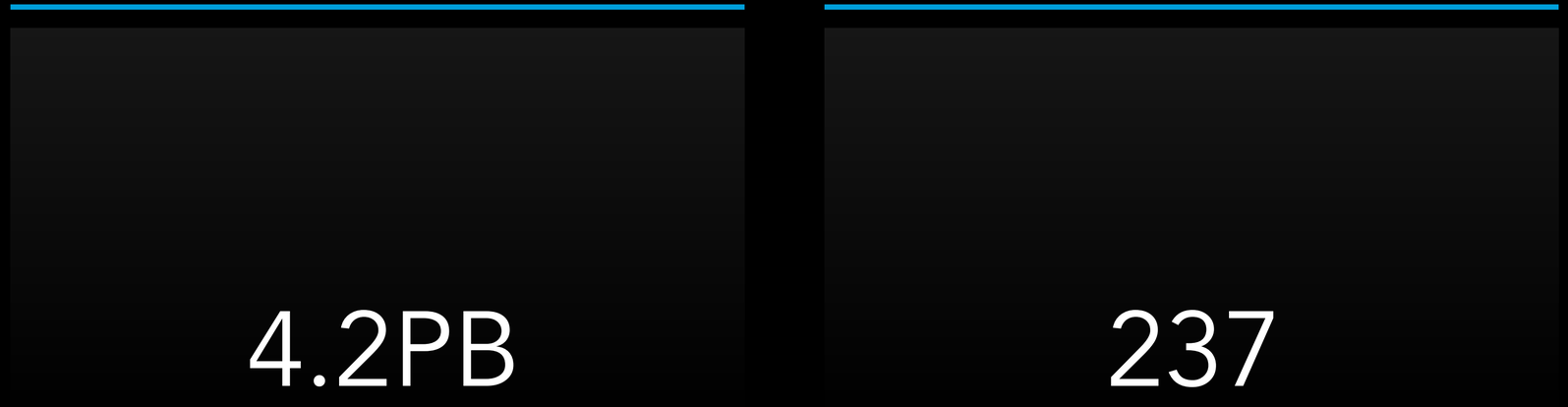


Stores LinkedIn's **Member Profile Data**.



- Distributed Document Store
- Strongly Consistent
- Provide features between NoSQL and RDBMS

Espresso Footprint



Storage Capacity

Databases

Espresso Capacity Planning



Organic Growth

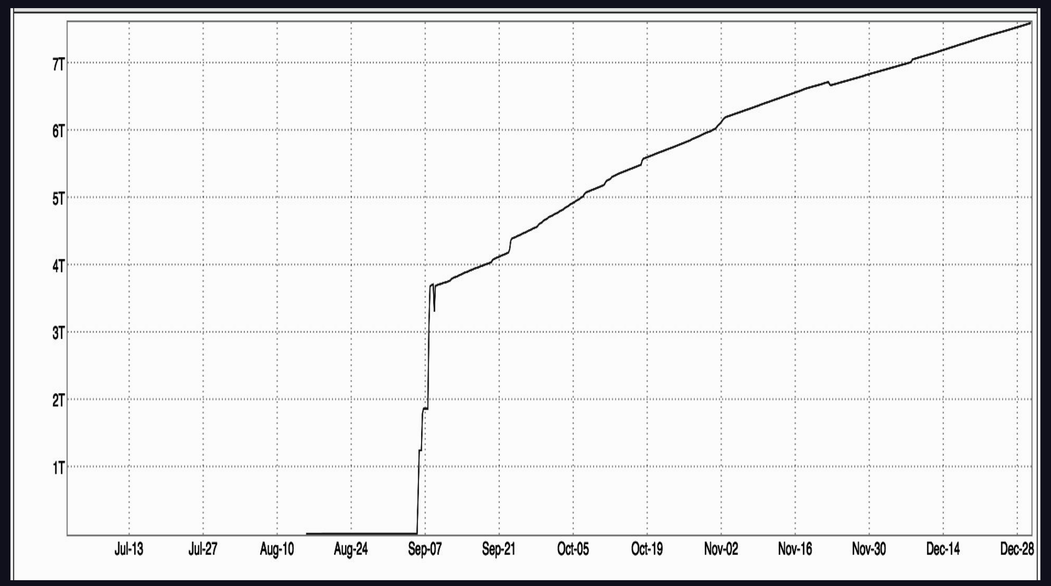


New Projects/ DBs



Unplanned Scenarios

Capacity model : Organic growth



InGraph (Disk Usage)

- Disk utilisation is not always linear
- Anomalies to be ignored
- Not max but percentile based calculation (~98th percentile)
- Quarterly projections are derived over weekly growth trend in previous quarter
- Buffer over projected capacity





- Distributed streaming platform
- Backbone of Linkedin’s data pipeline
- Metrics, member activity data, change data capture and much more

Kafka Footprint

2T

Kafka Events per Day

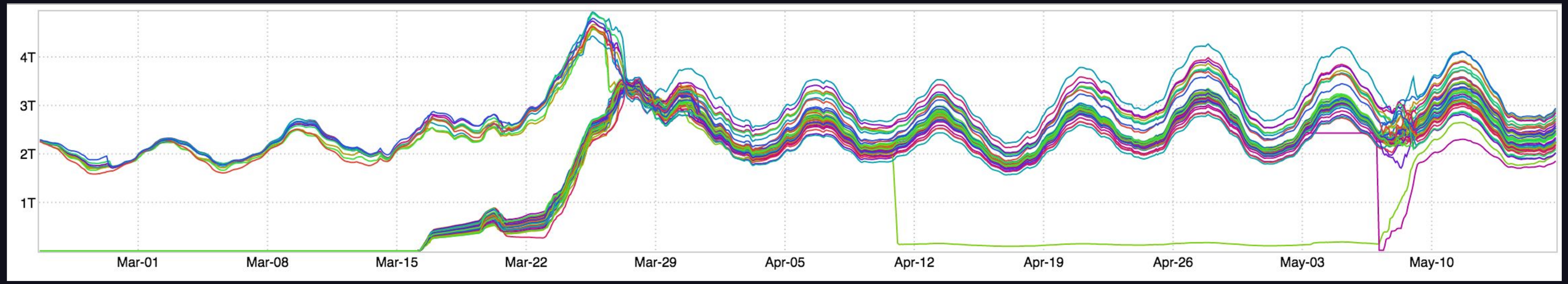
1600

Kafka Servers

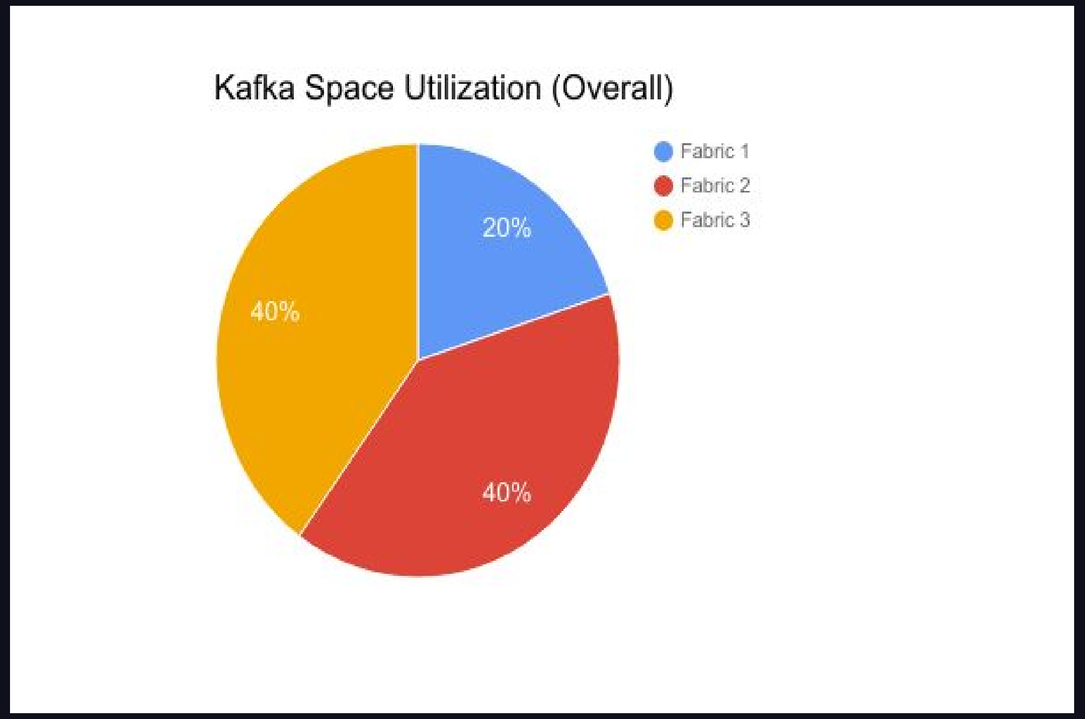
Kafka Capacity Planning

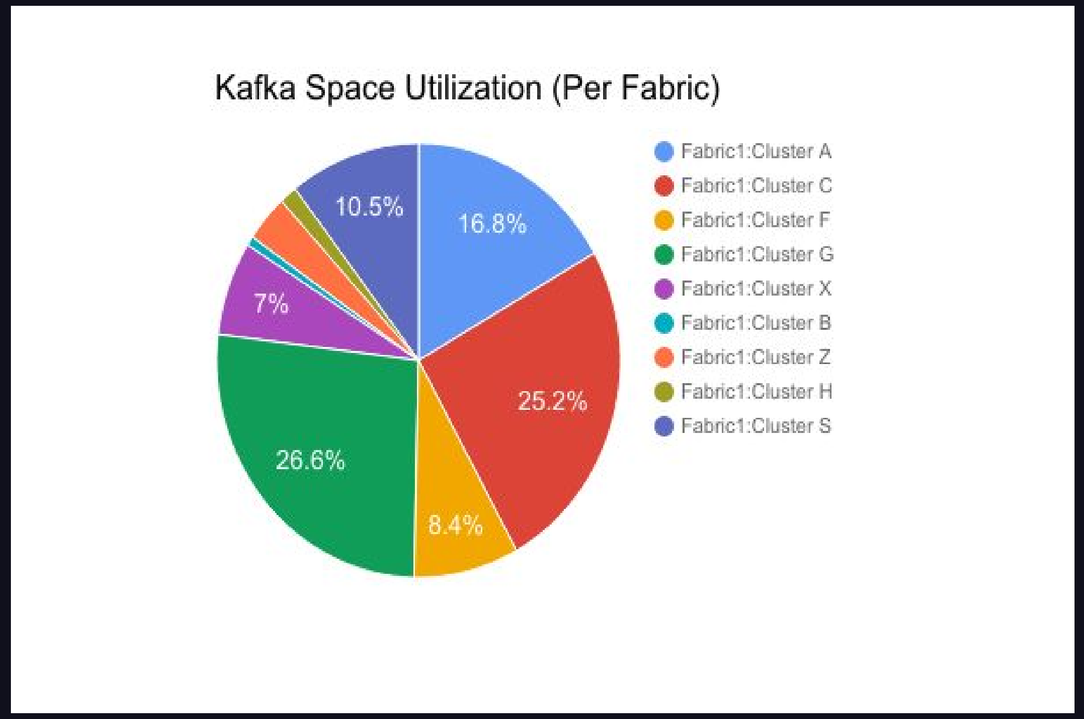
- Disk usage bound
- Cap disk utilisation at 60%
- Horizontally scalable

- Configurable retention
- Kafka rebalances to keep utilization under 40%

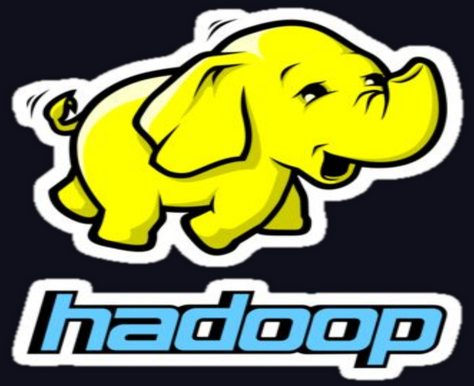


Kcap - Kafka Capacity Auditor









- Distributed storage and processing framework
- It can store any kind of data
- Write once read many

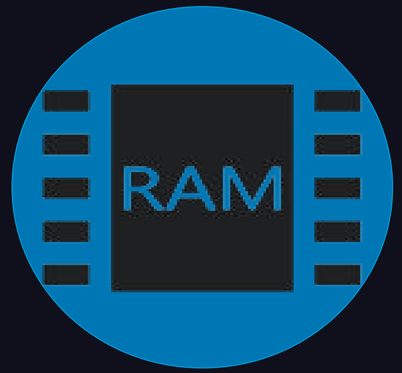
Hadoop Footprint

10K	170PB	145K	3.6K
Hadoop Nodes	HDFS Storage	YARN jobs/per day	Hadoop Users

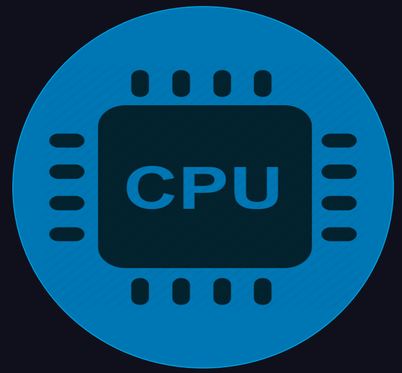
Hadoop Capacity Planning



Disk Utilization



Memory Utilization



CPU Utilization

Hadoop capacity management tools



Dr. Elephant



Rewinder



Sizr

Thank You

Q&A