



Aalto University
School of Science

Ingestion and Hadoop Case Studies

Hong-Linh Truong

Department of Computer Science

linh.truong@aalto.fi, <https://rdsea.github.io>

Case studies

- **Big Data Platform for monitoring in Slack**
- **Uber and Hadoop**
- **Goal: See how technologies we learn are used in real big platforms**
 - Key technologies: Kafka, Hadoop, Spark, ElasticSearch, Cloud storage
 - Techniques for data ingestion pipelines and integration
 - Real-world requirements

Slack

Read the paper and
analyze the case:
“Towards
Observability Data
Management at
Scale”

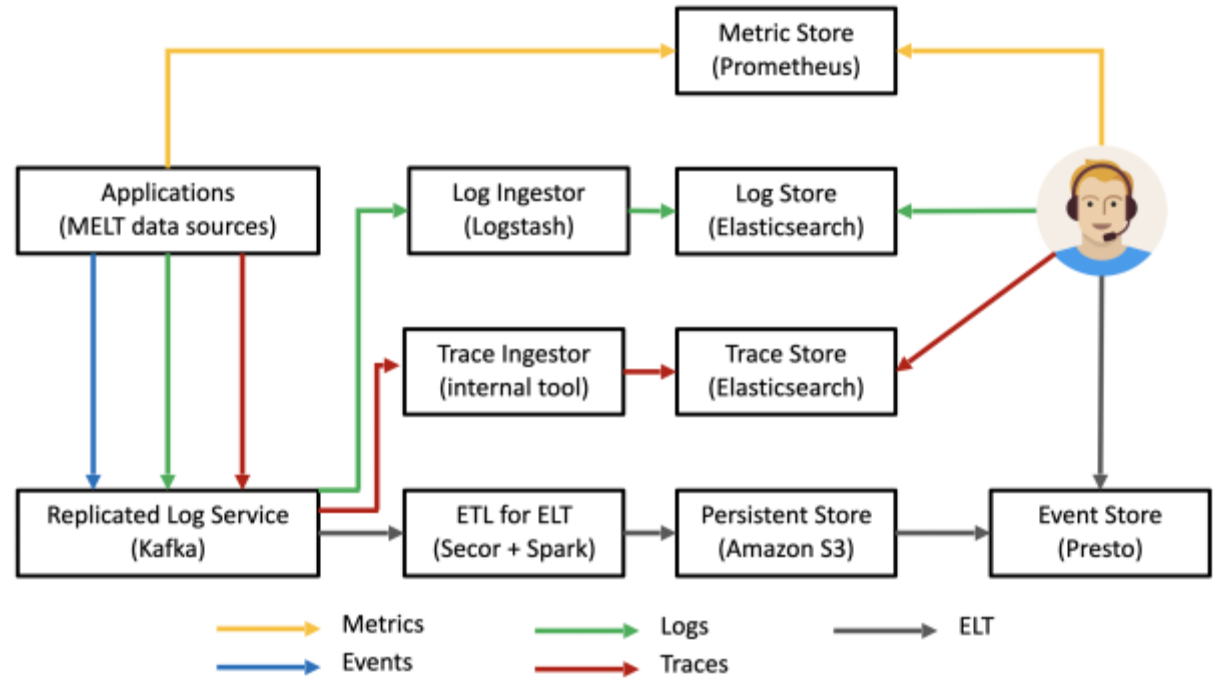


Figure source:

https://sigmodrecord.org/publications/sigmodRecord/2012/pdfs/05_Vision_Karumuri.pdf/

Uber case: study the role of Hadoop and Kafka

In this case study you are going to read:

<https://eng.uber.com/uber-big-data-platform/>

and examine the role of the Hadoop ecosystem and how Hadoop services work with other components.

Uber case: no Hadoop

Read the blog: <https://eng.uber.com/uber-big-data-platform/>

Generation 1 (2014-2015) - The beginning of Big Data at Uber

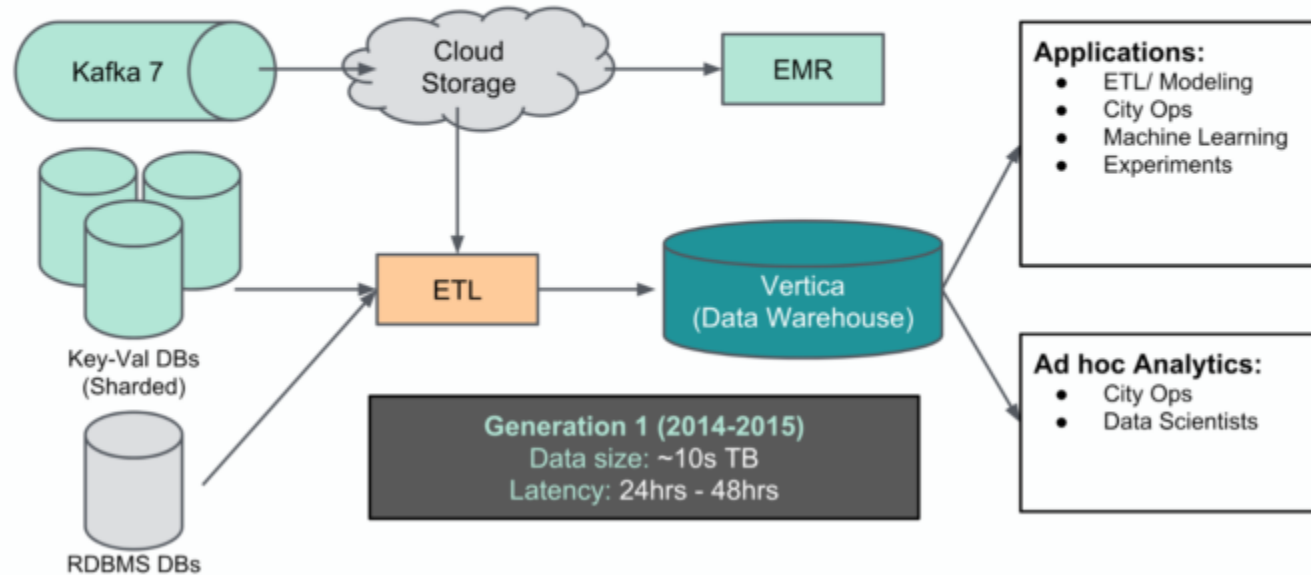


Figure source: <https://eng.uber.com/uber-big-data-platform/>

Uber case: with Hadoop

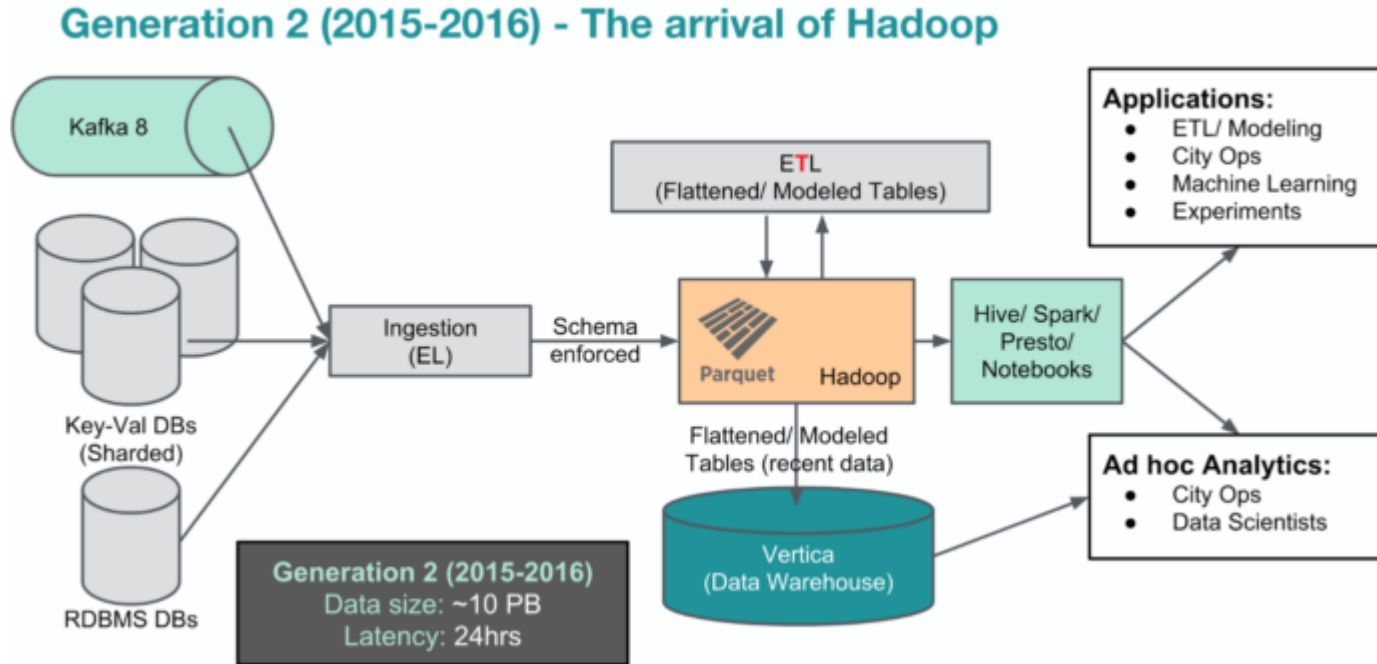


Figure source: <https://eng.uber.com/uber-big-data-platform/>

Uber case: with Hudi

Generation 3 (2017-present) - Let's rebuild for long term

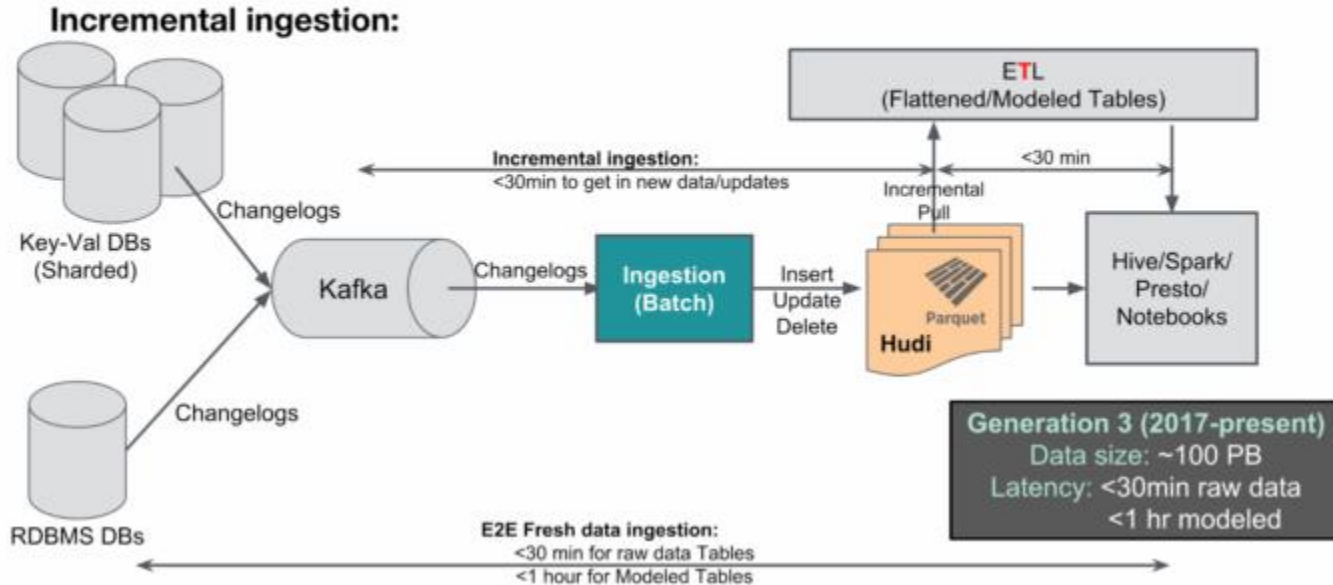


Figure source: <https://eng.uber.com/uber-big-data-platform/>
Read Hudi: <https://eng.uber.com/hoodie/>