State of
Prezi Data Platform

Data Platform

Define, Ingest, Store, Process, Orchestrate
User Needs

Architectures

Data Warehouse
Table storage + Compute

Data Lake
[illegible]

Data Lakehouse
[illegible]

Introduction

Exploring the Prezi Data Platform, its architecture as a Data Lakehouse, and the key components.

- What is a data platform and why we need it
- Different Architectures and DataLakehouse
- Components

00

Definition

Events Definition

- Logmanagement (Django)
- New schema in Schema Registry
- New topic in Kafka
- Logging Libraries (all languages)

01

Ingestion

Efficient, real-time data ingestion ensures seamless data flow into the Data Lakehouse

- [illegible] + Filebeat
- Logstash
- Kafka -
- Kafka Connect

02

Storage

The robust storage of the Data Lakehouse accommodates structured, unstructured, and semi-structured data, enabling scalable and efficient data management.

S3 + Parquet + Hive

03

Processing

In the Data Lakehouse architecture, processing is offloaded to a separate component. In our infra we use Spark and Trino

[illegible]

04

Orchestration

The numerous jobs we have, are coordinated/orchestrated by a component called [illegible]. It's responsibility is to run the correct job at the correct time

Conclusion

To sum it up:

- Prezi Data Platform is founded on stable/state-of-the-art techs and architecture decision
- It allows the introduction of cutting edge technology, as it is pluggable and modular
- It's constantly evolving

Trino Summit 2024

Thanks!

@vincenzocassaro

Introduction

Exploring the Prezi Data Platform, its architecture as a Data Lakehouse, and the key components.

- What is a data platform and why we need it
- Different Architectures and DataLakehouse
- Components

Vincenzo Cassaro

Data Engineer @Prezi

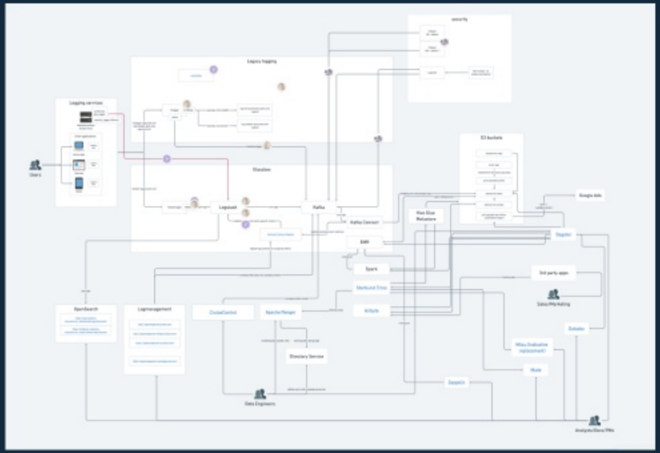
Data Infrastructure Team

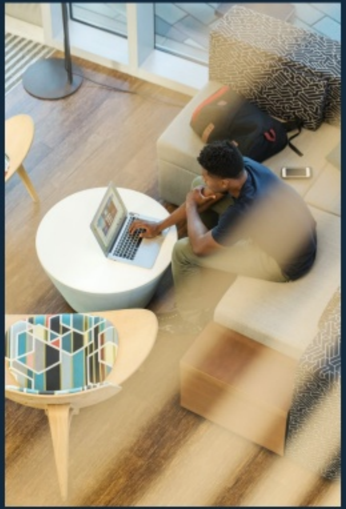
From Italy🇮🇹

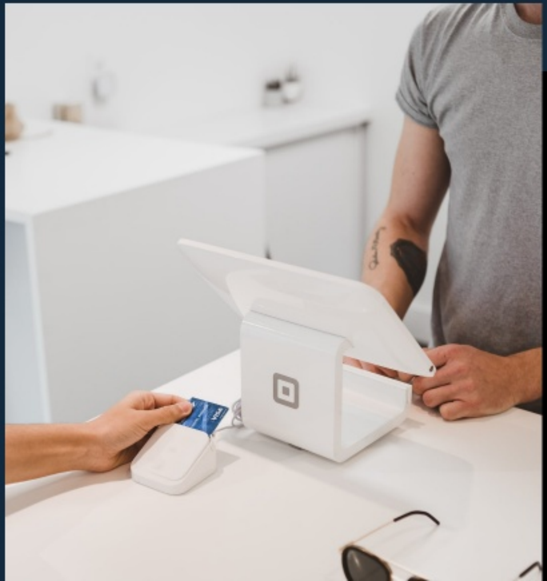
Data Platform

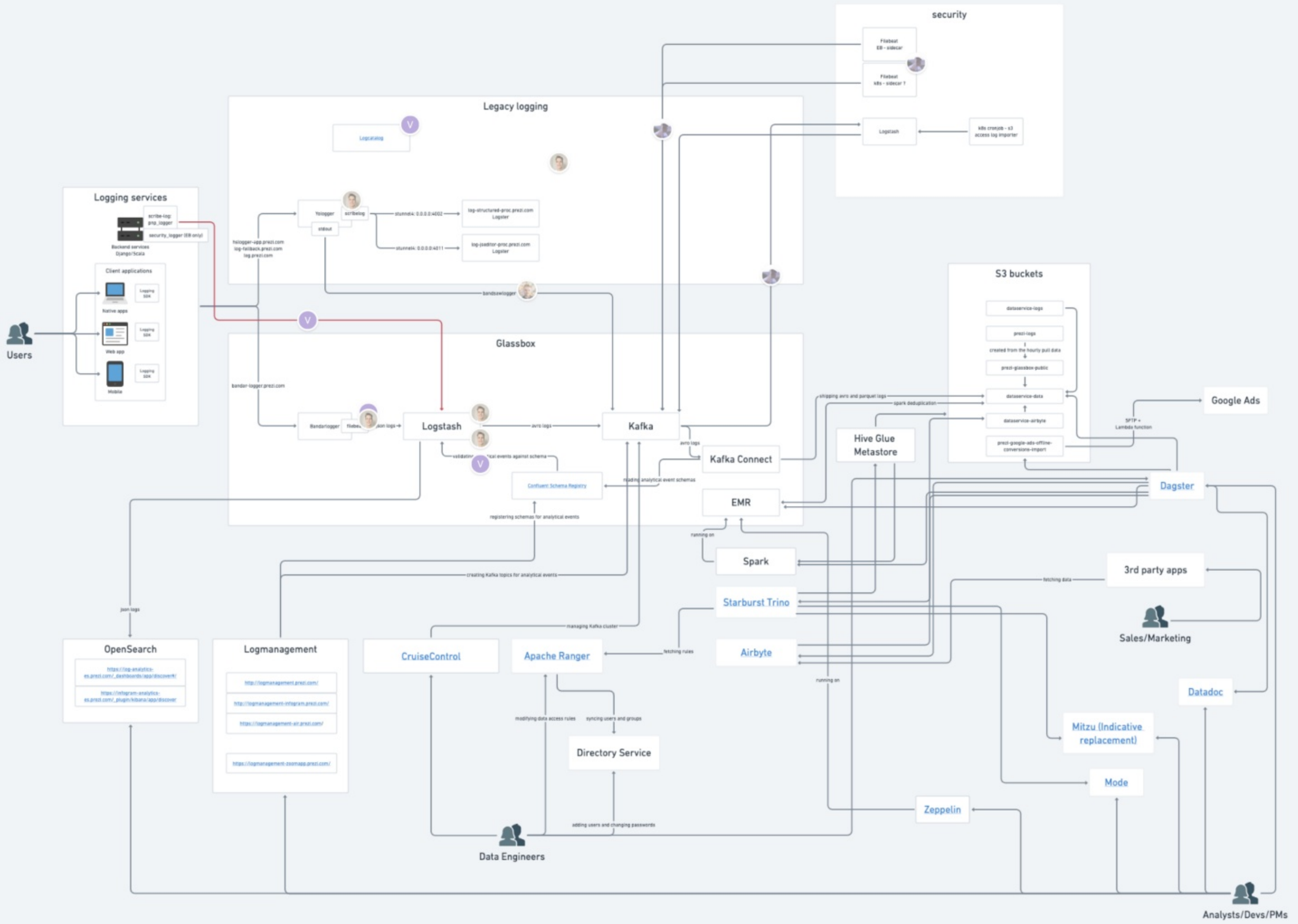
Define, Ingest, Store, Process, Orchestrate

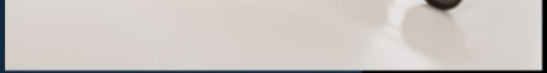
User Events

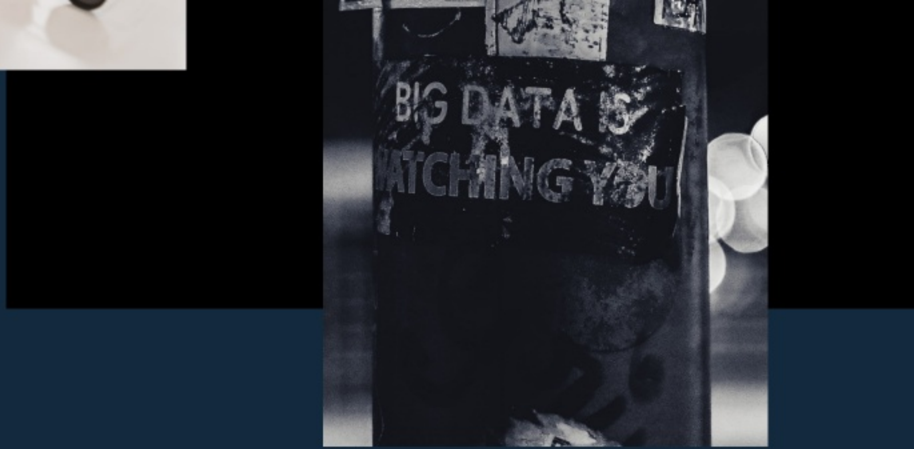







security
Filebeat
EB - sidecar
Filebeat
k8s - sidecar ?
Logstash
k8s cronjob - s3
access log importer
Legacy logging
Logcatalog
Logging services
scribe-log:
pnp_logger
security_logger (EB only)
Backend services
Django/Scala
Client applications
Native apps
Logging SDK
Web app
Logging SDK
Mobile
Logging SDK
Users
Yologger
scribelog
stdout
stunnel4: 0.0.0.0:4002
log-structured-proc.prezi.com
Logster
stunnel4: 0.0.0.0:4011
log-jseditor-proc.prezi.com
Logster
yologger-app.prezi.com
log-fallback.prezi.com
log.prezi.com
bandsawlogger
Glassbox
bandar-logger.prezi.com
Bandarlogger
filebe...
...son logs
Logstash
avro logs
Kafka
validatin... ...ical events against schema
Confluent Schema Registry
reading analytical event schemas
registering schemas for analytical events
avro logs
Kafka Connect
EMR
running on
Spark
Starburst Trino
Airbyte
Hive Glue
Metastore
S3 buckets
dataservice-logs
prezi-logs
created from the hourly pull data
prezi-glassbox-public
dataservice-data
dataservice-airbyte
prezi-google-ads-offline-conversions-import
shipping avro and parquet logs
spark deduplication
SFTP +
Lambda function
Google Ads
Dagster
3rd party apps
fetching data
Sales/Marketing
Datadoc
Mitzu (Indicative replacement)
Mode
Zeppelin
running on
json logs
OpenSearch
https://log-analytics-es.prezi.com/_dashboards/app/discover#/
https://infogram-analytics-es.prezi.com/_plugin/kibana/app/discover
Logmanagement
http://logmanagement.prezi.com/
http://logmanagement-infogram.prezi.com/
https://logmanagement-air.prezi.com/
https://logmanagement-zoomapp.prezi.com/
creating Kafka topics for analytical events
managing Kafka cluster
CruiseControl
Apache Ranger
fetching rules
modifying data access rules
syncing users and groups
Directory Service
adding users and changing passwords
Data Engineers
Analysts/Devs/PMs





Architectures

Data Warehouse
Table storage + Compute

Data Lake
Multiple format (messy) storage

Data LakeHouse
Best of Both World (free file format + Table Api)

Architectures

Data Warehouse
Table storage + Compute

Data Lake
Multiple format (messy) storage

Data LakeHouse
Best of Both World (free file format + Table Api)

State of
Prezi Data Platform

Introduction

Exploring the Prezi Data Platform, its architecture as a Data Lakehouse, and the key components.

- What is a data platform and why we need it
- Different Architectures and DataLakehouse
- Components

Data Platform

Define. Ingest. Store. Process. Orchestrate. User Needs

Architectures

Data Warehouse
Table storage + Compute

Data Lake
Multiple format storage

Data Lakehouse
[illegible]

00

Definition

Events Definition

- Logmanagement (Django)
- New schema in Schema Registry
- New topic in Kafka
- Logging Libraries (all languages)

01

Ingestion

Efficient, real-time data ingestion ensures seamless data flow into the Data Lakehouse

- [illegible] + Filebeat
- Logstash
- Kafka
- Kafka Connect

02

Storage

The robust storage of the Data Lakehouse accommodates structured, unstructured, and semi-structured data, enabling scalable and efficient data management.

S3 + Parquet + Hive

03

Processing

[illegible]

[illegible]

04

Orchestration

[illegible]

Conclusion

To sum it up:

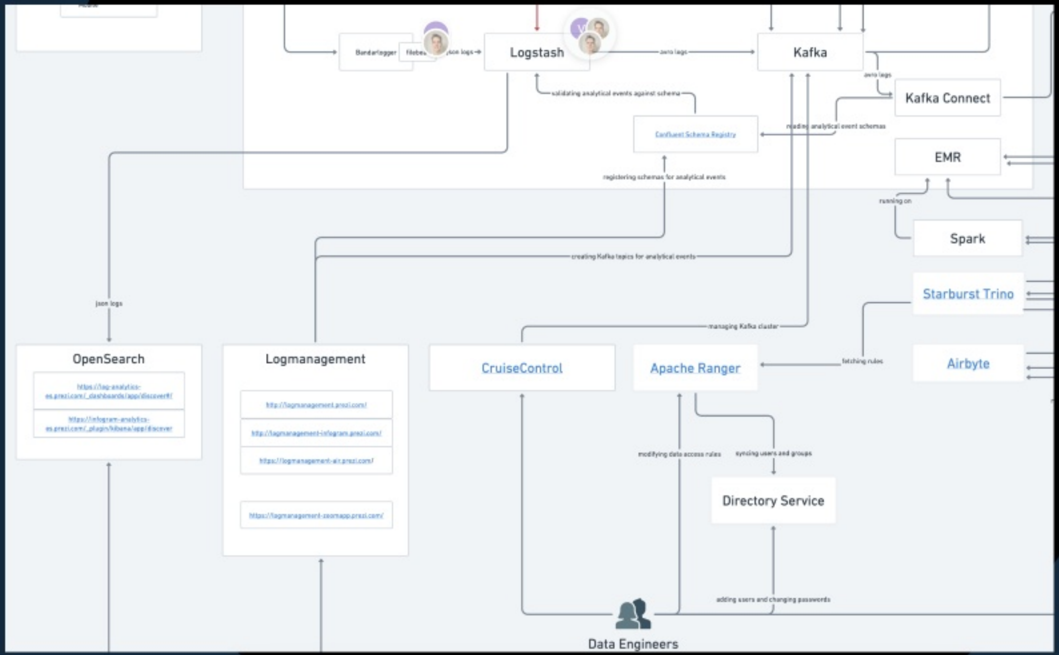
- Prezi Data Platform is founded on stable/state-of-the-art techs and architecture decision
- It allows the introduction of cutting edge technology, as it is pluggable and modular
- It's constantly evolving

Trino Summit 2024

Thanks!

@vincenzocassaro

00

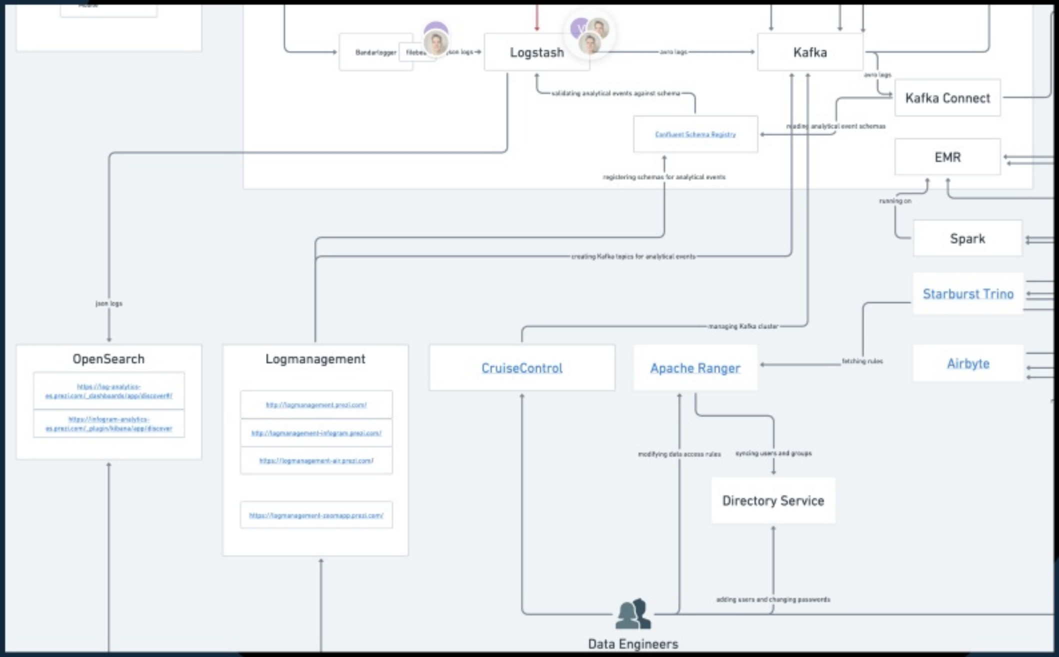

Bandarlogger
filebe
json logs
Logstash
avro logs
Kafka
avro logs
Kafka Connect
validating analytical events against schema
Confluent Schema Registry
reading analytical event schemas
EMR
registering schemas for analytical events
running on
Spark
creating Kafka topics for analytical events
Starburst Trino
json logs
managing Kafka cluster
OpenSearch
https://log-analytics-es.prezi.com/_dashboards/app/discover#/
https://infogram-analytics-es.prezi.com/_plugin/kibana/app/discover
Logmanagement
http://logmanagement.prezi.com/
http://logmanagement-infogram.prezi.com/
https://logmanagement-air.prezi.com/
https://logmanagement-zoomapp.prezi.com/
CruiseControl
Apache Ranger
fetching rules
Airbyte
modifying data access rules
syncing users and groups
Directory Service
adding users and changing passwords
Data Engineers

Definition

Events Definition

- Logmanagement (Django)
- New schema in Schema Registry
- New topic in Kafka
- Logging Libraries (all languages)

00

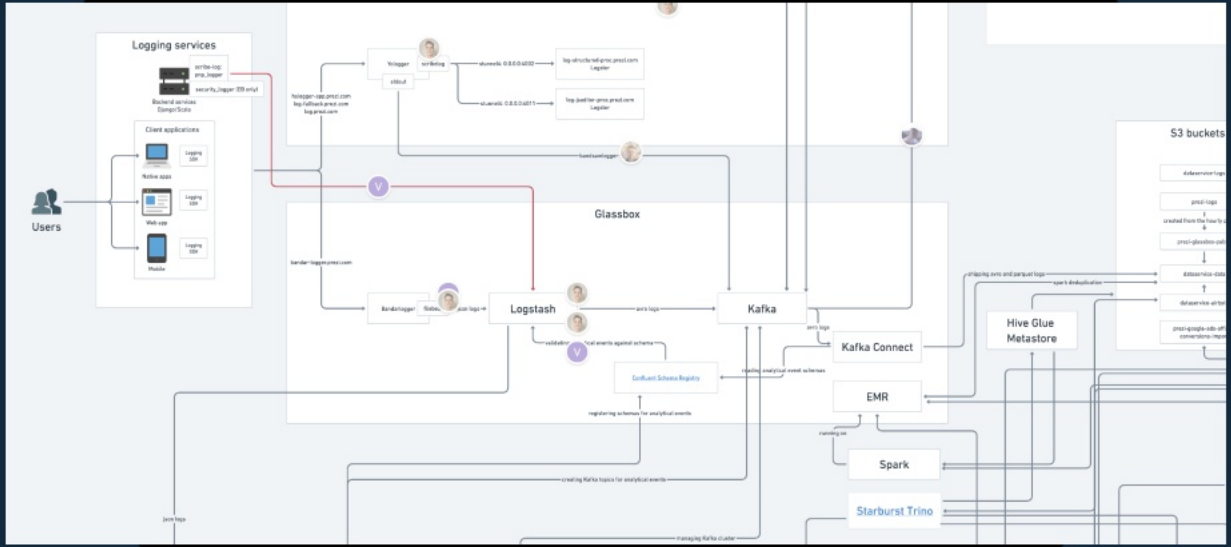

Bandarlogger
filebe
json logs
Logstash
avro logs
Kafka
avro logs
Kafka Connect
validating analytical events against schema
Confluent Schema Registry
reading analytical event schemas
EMR
registering schemas for analytical events
running on
Spark
creating Kafka topics for analytical events
Starburst Trino
json logs
managing Kafka cluster
OpenSearch
https://log-analytics-es.prezi.com/_dashboards/app/discover#/
https://infogram-analytics-es.prezi.com/_plugin/kibana/app/discover
Logmanagement
http://logmanagement.prezi.com/
http://logmanagement-infogram.prezi.com/
https://logmanagement-air.prezi.com/
https://logmanagement-zoomapp.prezi.com/
CruiseControl
Apache Ranger
fetching rules
Airbyte
modifying data access rules
syncing users and groups
Directory Service
adding users and changing passwords
Data Engineers

Definition

Events Definition

- Logmanagement (Django)
- New schema in Schema Registry
- New topic in Kafka
- Logging Libraries (all languages)

01



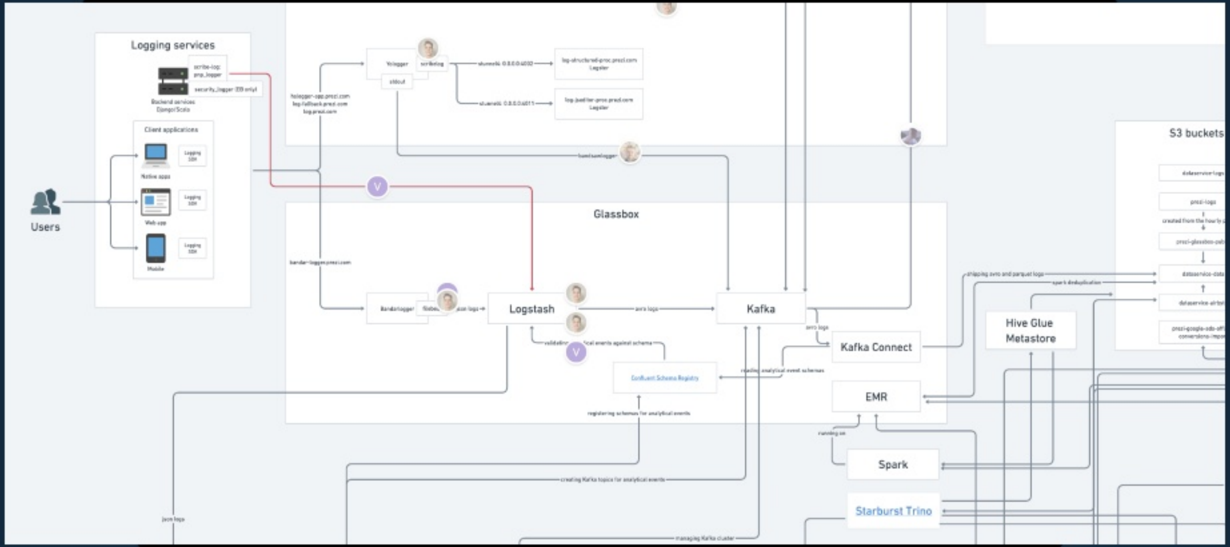
Ingestion

Efficient, real-time data ingestion ensures seamless data flow into the Data Lakehouse

- Bandarlogger + Filebeat
- Logstash
- Kafka
- Kafka Connect

kafka
16 Brokers
1.5 TB a day
~3000 topics/tables

01



Ingestion

Efficient, real-time data ingestion ensures seamless data flow into the Data Lakehouse

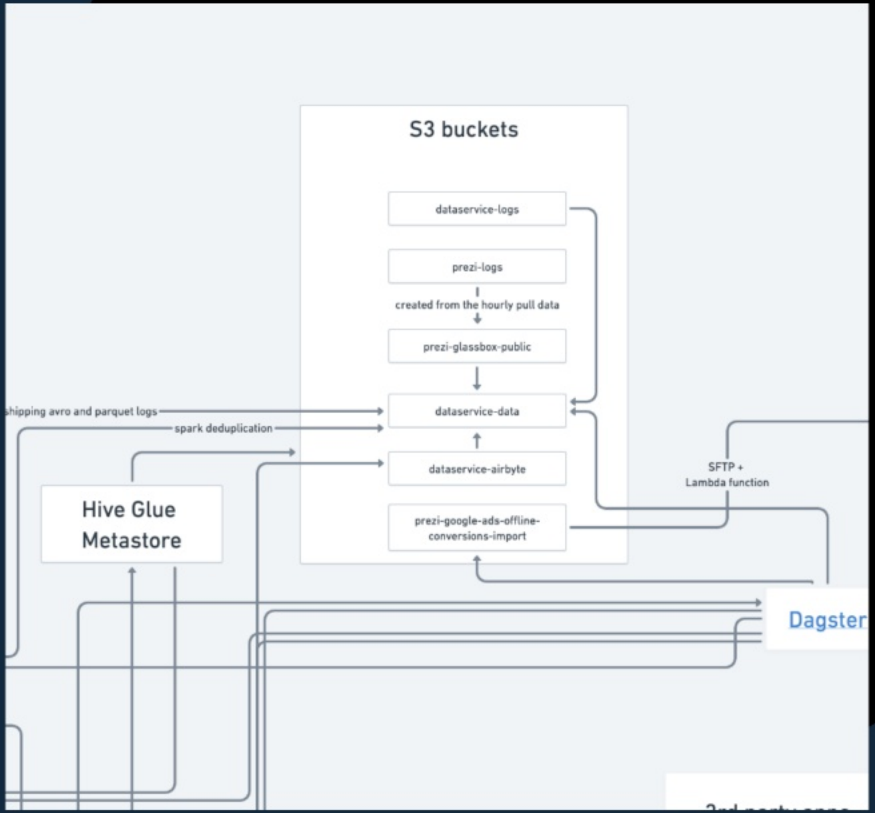
- Bandarlogger + Filebeat
- Logstash
- Kafka kafka 16 Brokers 1.5 TB a day ~3000 topics/tasks
- Kafka Connect

"On Premise" -> MSK, Gobblin

kafka

16 Brokers
1.5 TB a day
~3600 topics/tables

02

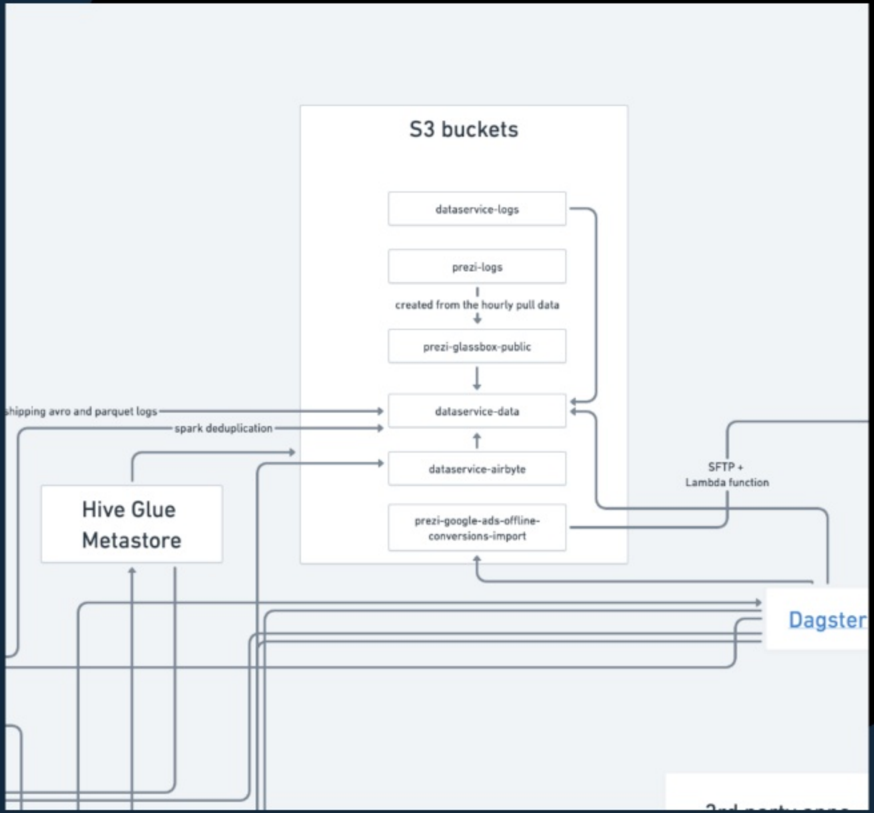


Storage

The robust storage of the Data Lakehouse accommodates structured, unstructured, and semi-structured data, enabling scalable and efficient data management.

S3 + Parquet + Hive

02



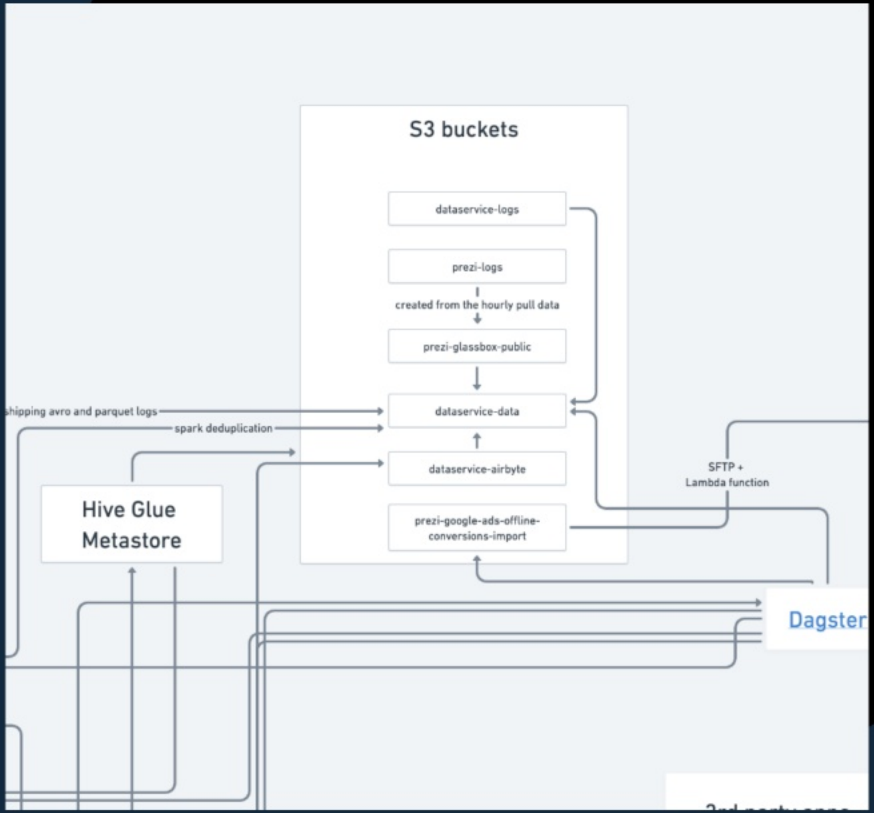
Storage

The robust storage of the Data Lakehouse accommodates structured, unstructured, and semi-structured data, enabling scalable and efficient data management.

S3 + Parquet + Hive

Hive Metastore -> Glue

02



Storage

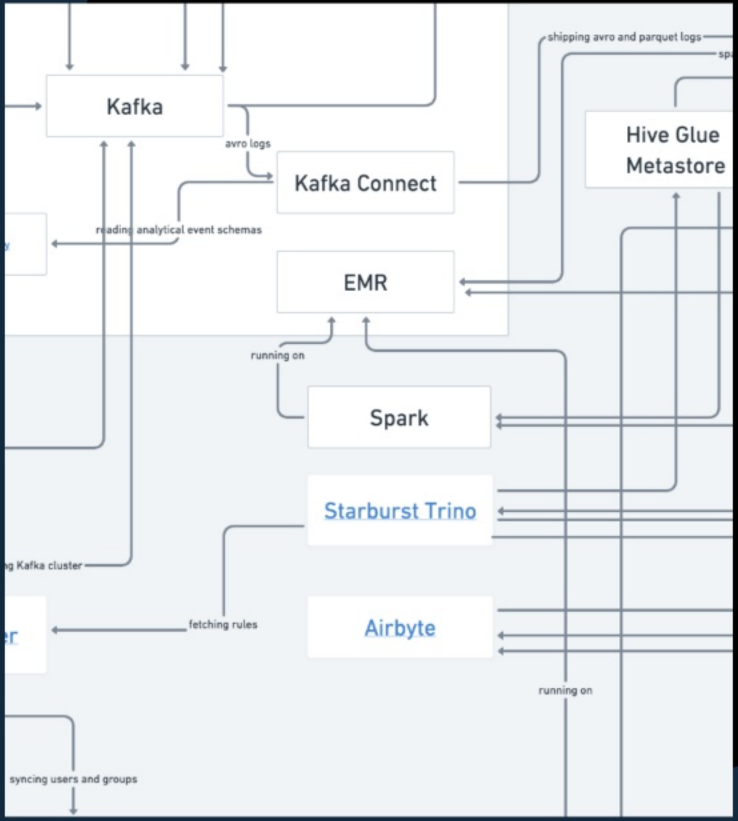
The robust storage of the Data Lakehouse accommodates structured, unstructured, and semi-structured data, enabling scalable and efficient data management.

S3 + Parquet + Hive

Hive Metastore -> Glue

Iceberg

03

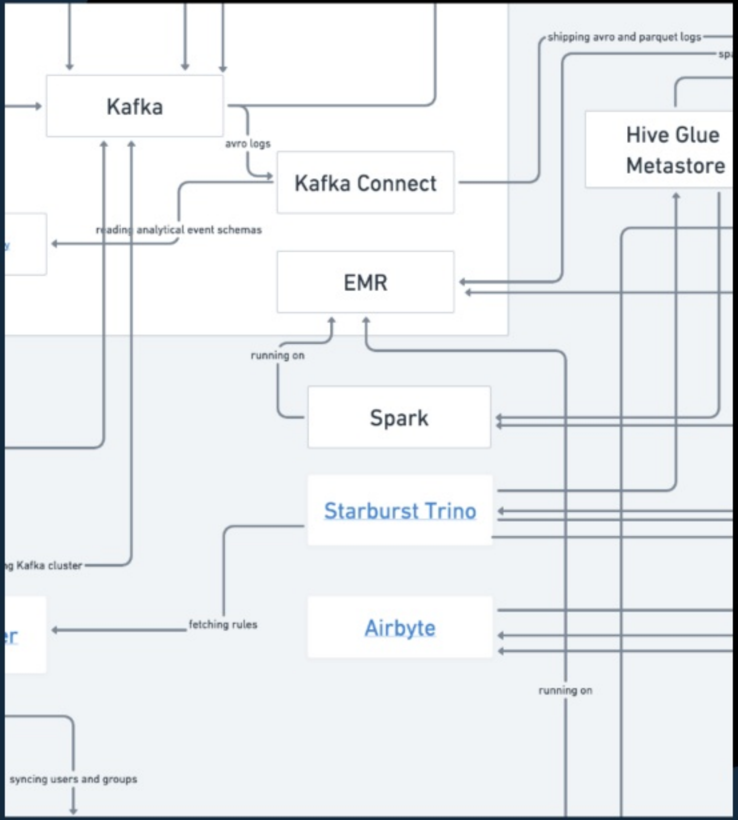


Processing

I the Data Lakehouse Architecture, processing is offloaded to a separate component. In our infra we use Spark and Trino

7 x r5.2xlarge

03



Processing

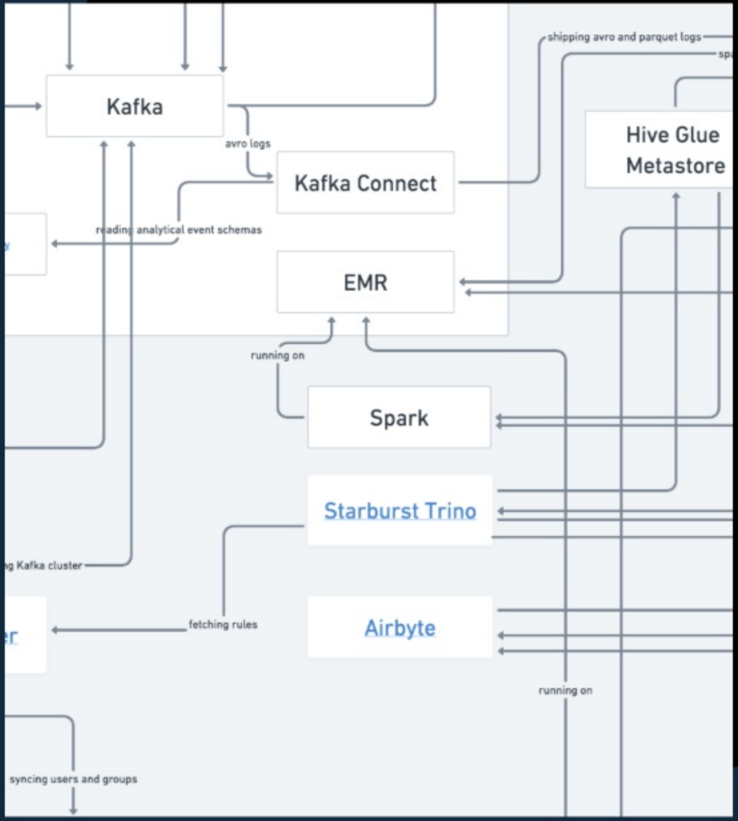
I the Data Lakehouse Architecture, processing is offloaded to a separate component. In our infra we use Spark and Trino

7 x r5.2xlarge

Spark 3.5

03

Processing



I the Data Lakehouse Architecture, processing is offloaded to a separate component. In our infra we use Spark and Trino

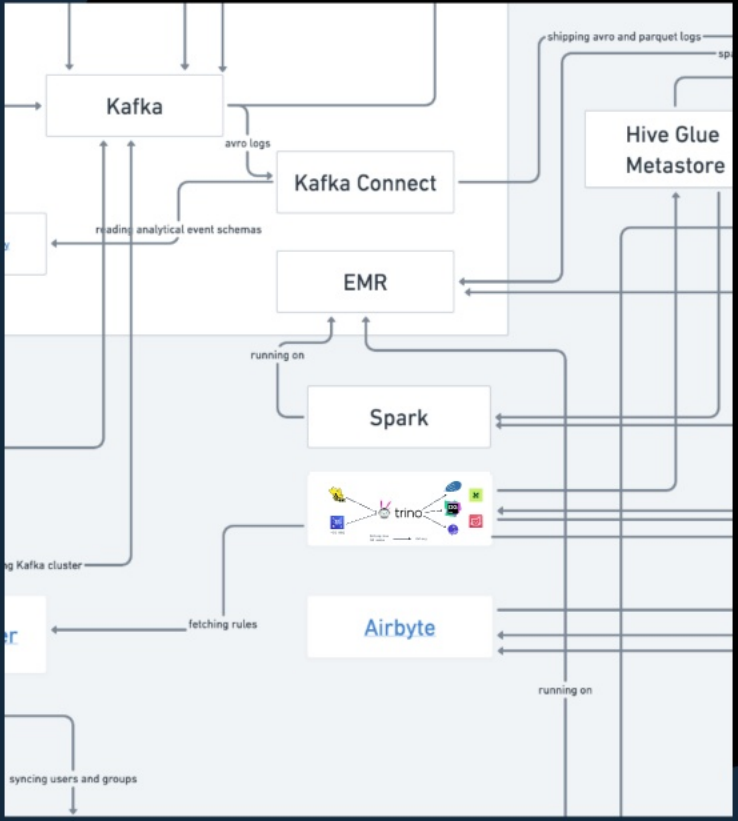
7 x r5.2xlarge

Spark 3.5

Scalability and saas

03

Processing

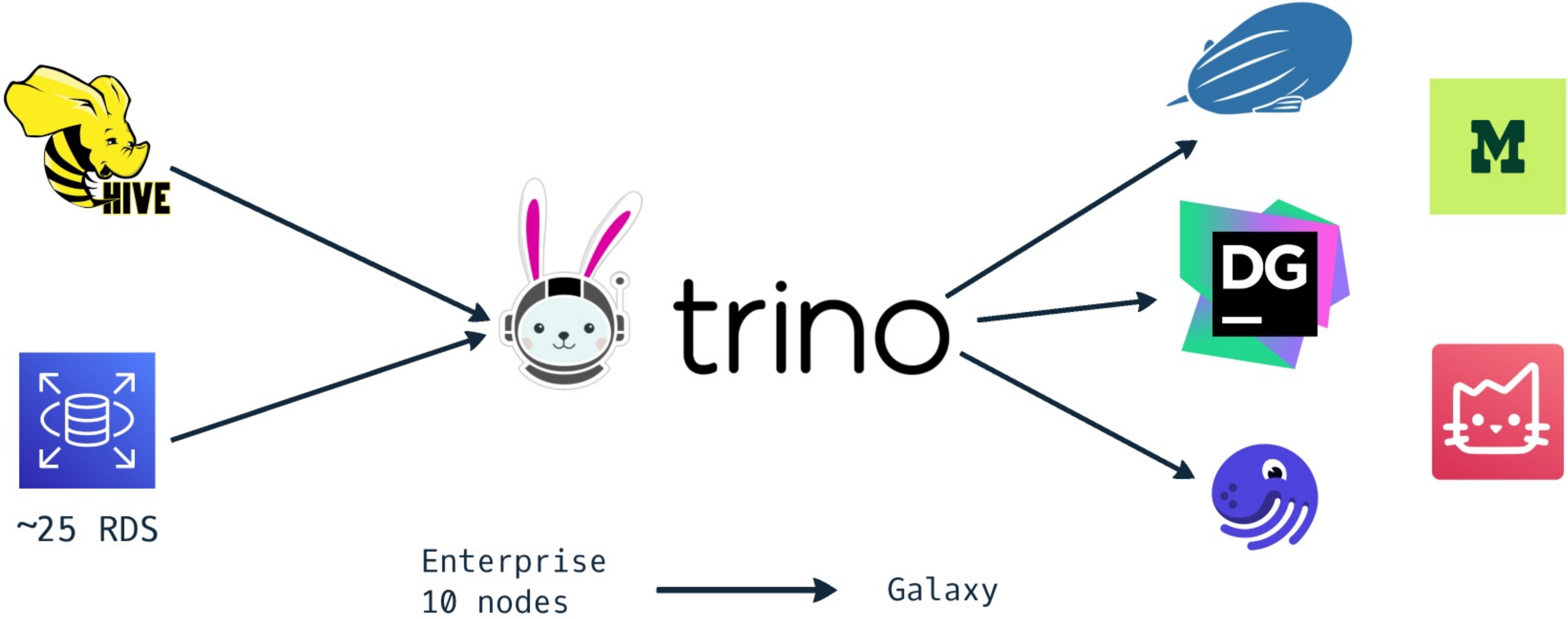


I the Data Lakehouse Architecture, processing is offloaded to a separate component. In our infra we use Spark and Trino

7 x r5.2xlarge

Spark 3.5

Scalability and saas



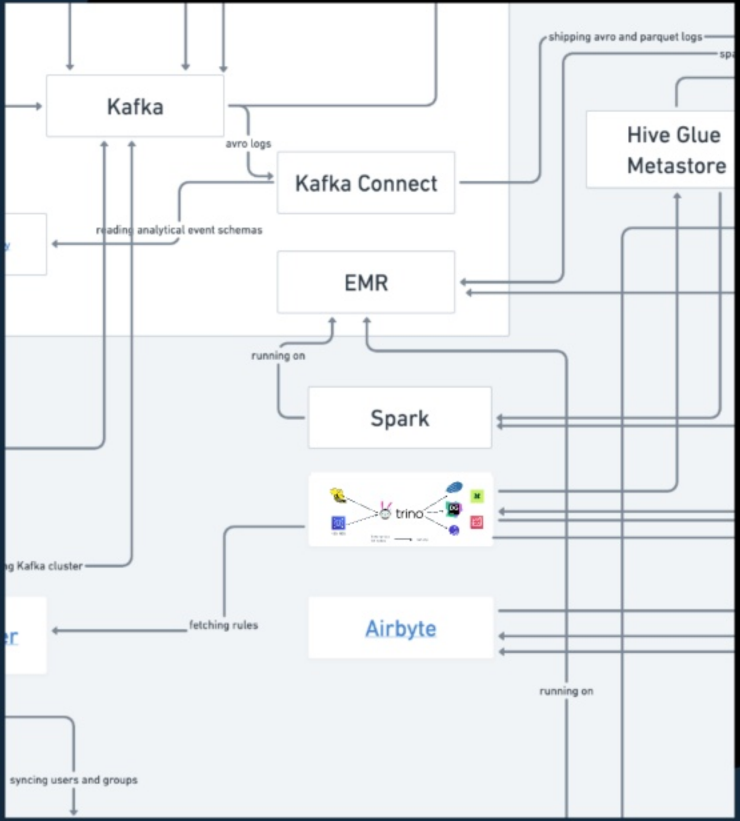
~25 RDS

Enterprise
10 nodes

Galaxy

03

Processing



I the Data Lakehouse Architecture, processing is offloaded to a separate component. In our infra we use Spark and Trino

7 x r5.2xlarge

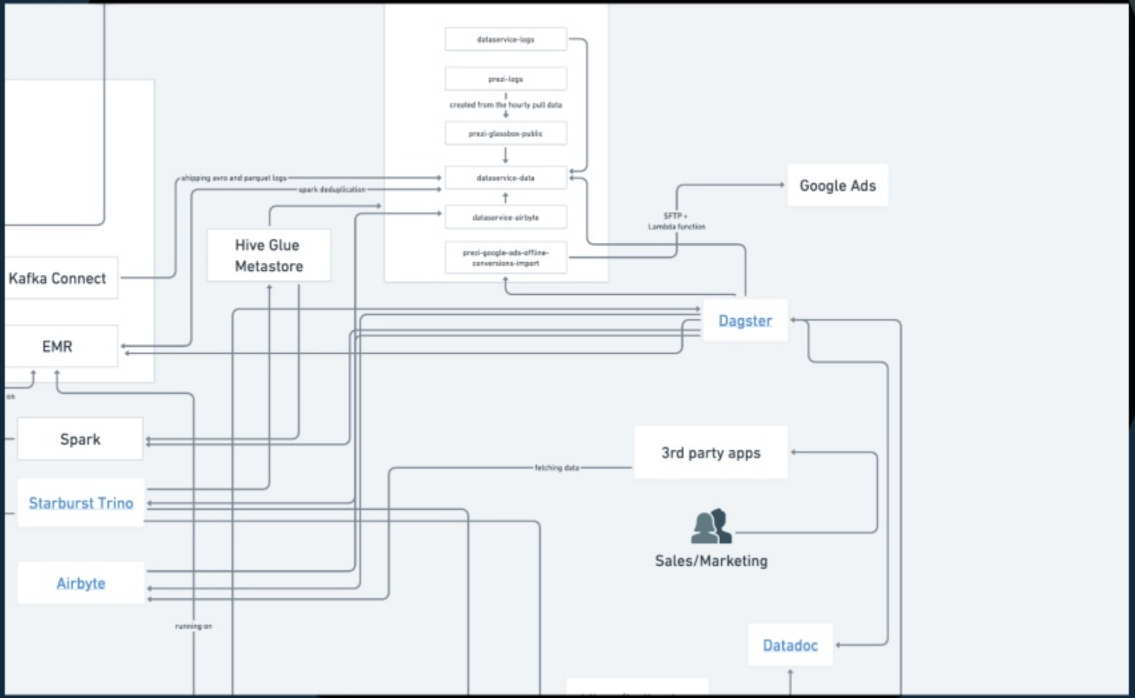
Spark 3.5

Scalability and saas

04

Orchestration

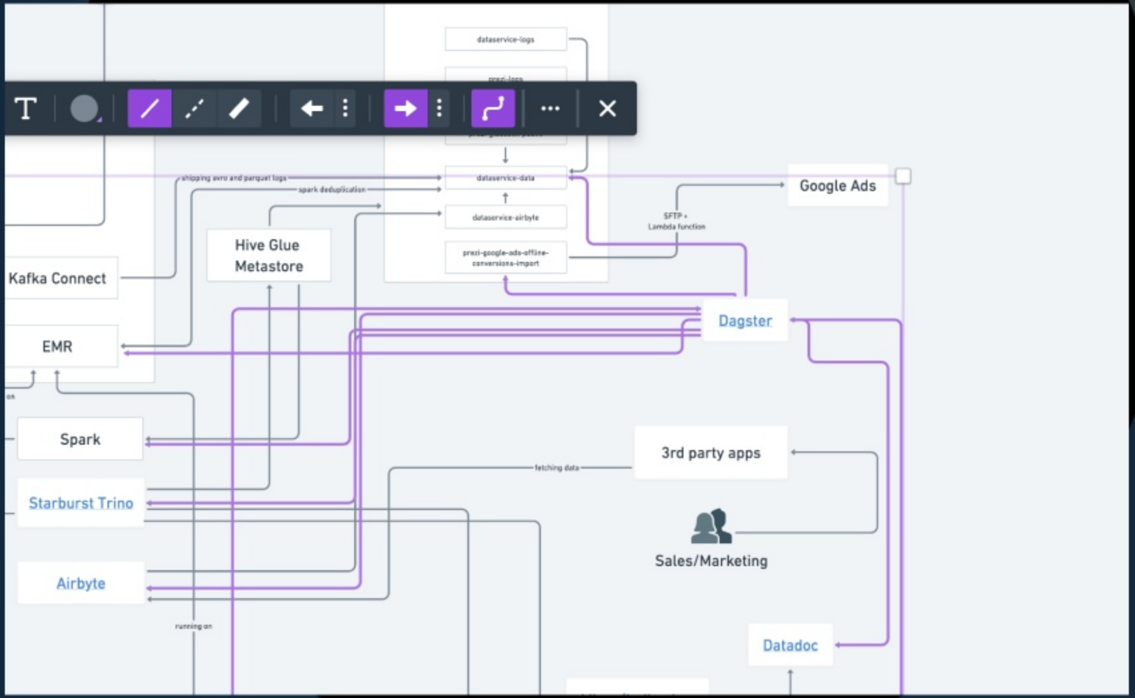
The numerous jobs we have, are coordinated/orchestrated by a component called Dagster. It's responsibility is to run the correct job at the correct time



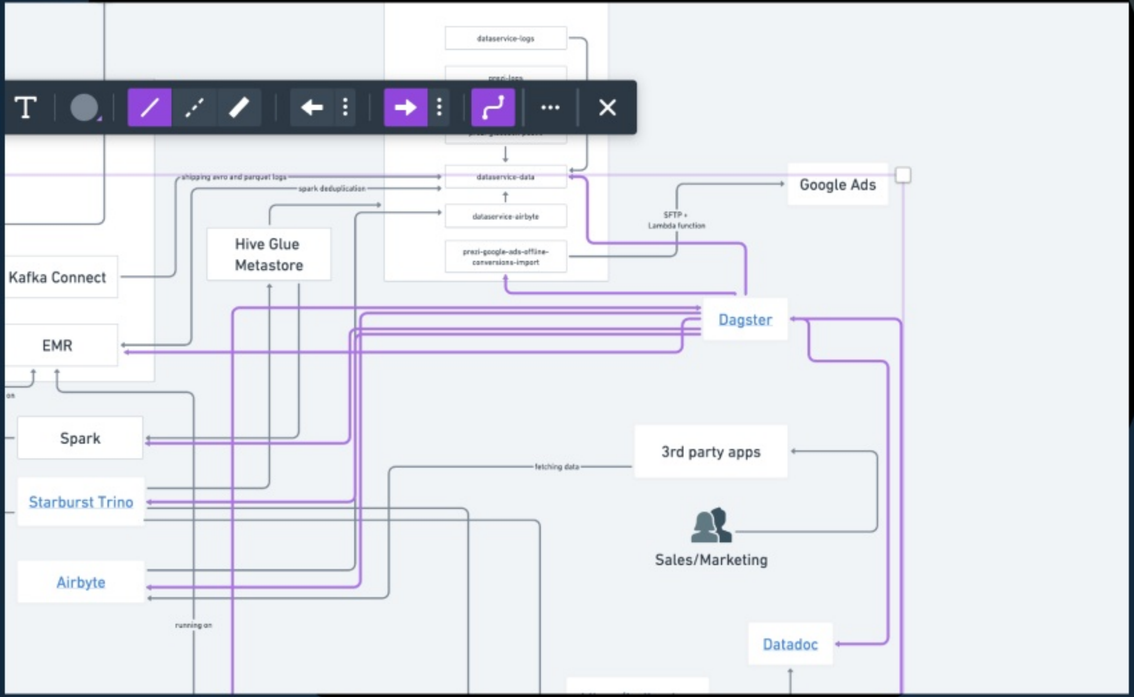
04

Orchestration

The numerous jobs we have, are coordinated/orchestrated by a component called Dagster. It's responsibility is to run the correct job at the correct time



04



Orchestration

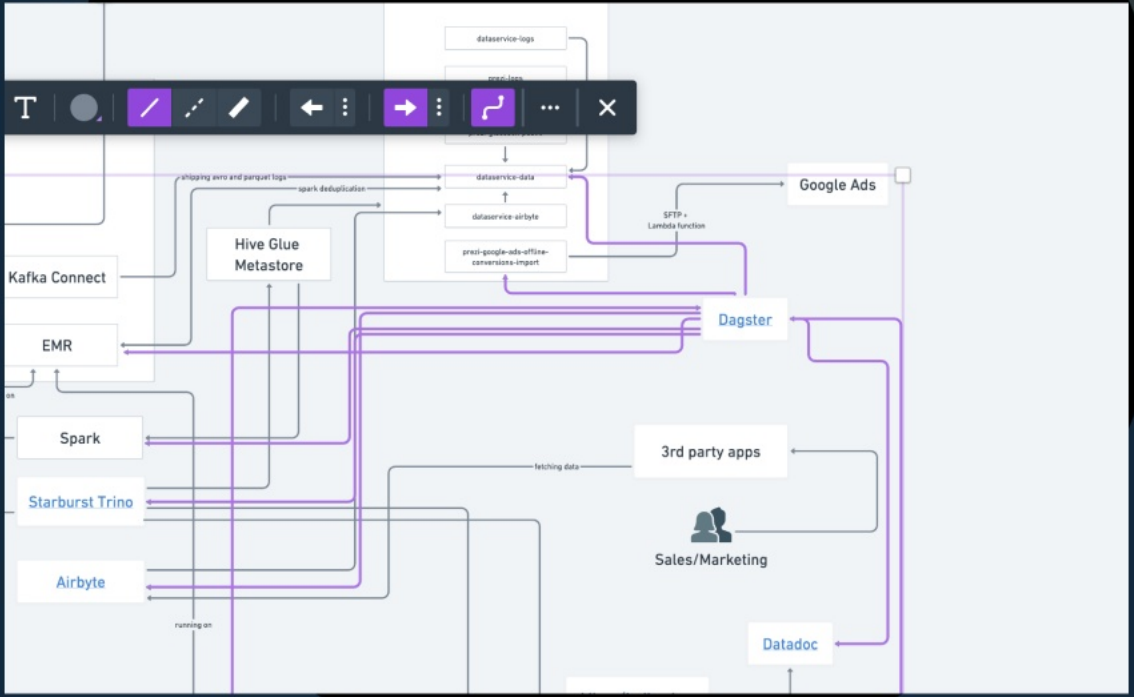
The numerous jobs we have, are coordinated/orchestrated by a component called Dagster. It's responsibility is to run the correct job at the correct time

More Flowkeeperless

04

Orchestration

The numerous jobs we have, are coordinated/orchestrated by a component called Dagster. It's responsibility is to run the correct job at the correct time



More Flowkeeperless

Hybrid Saas

Data Platform

Define. Ingest. Store. Process. Orchestrate [illegible]

Architectures

Data Warehouse
Table storage + Compute

Data Lake
[illegible] storage

Data Lakehouse
[illegible]

State of
Prezi Data Platform

Introduction

Exploring the Prezi Data Platform, its architecture as a Data Lakehouse, and the key components.

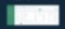
- What is a data platform and why we need it
- Different Architectures and DataLakehouse
- Components

00

Definition

Events Definition

- Logmanagement (Django)
- New schema in Schema Registry
- New topic in Kafka
- Logging Libraries (all languages)



01

Ingestion

Efficient, real-time data ingestion ensures seamless data flow into the Data Lakehouse

- Random logger + FileBeat
- Logstash
- Kafka -
- Kafka Connect

On Premise -> AWS [illegible]

02

Storage

The robust storage of the Data Lakehouse accommodates structured, unstructured, and semi-structured data, enabling scalable and efficient data management.

S3 + Parquet + Hive

Hive Metastore -> Glue

Iceberg

03

Processing

In the Data Lakehouse architecture, processing is [illegible] to a specific component. In our infra we use Spark and Trino

2 x [illegible]

Spark 3.5

Scalability and [illegible]

04

Orchestration

The numerous jobs we have, are coordinated/orchestrated by a component called Dagster. It's responsibility is to run the correct job at the correct time

More [illegible]

Hybrid [illegible]

Conclusion

To sum it up:
- Prezi Data Platform is founded on stable/state-of-the-art techs and architecture decision
- It allows the introduction of cutting edge technology, as it is pluggable and modular
- It's constantly evolving

Trino Summit 2024

Thanks!

@vincenzocassaro

Conclusion

To sum it up:

- Prezi Data Platform is founded on stable/state-of-the-art techs and architecture decision
- It allows the introduction of cutting edge technology, as it is pluggable and modular
- It's constantly evolving

Trino Summit 2024

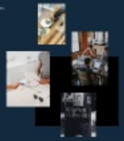
@vincenzocassaro

Thanks!

State of Prezi Data Platform

Data Platform

[illegible]



Architecture

Data Warehouse
[illegible]

Data Lake
[illegible]

Data Lakehouse
[illegible]

Introduction

Exploring the Prezi Data Platform, its architecture as a Data Lakehouse, and the key components.

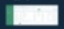
- What is a data platform and why we need it
- Different Architectures and DataLakehouse
- Components

00

Definition

Events Definition

- Logmanagement (Django)
- New schema in Schema Registry
- New topic in Kafka
- Logging Libraries (all languages)



01

Ingestion

Efficient, real-time data ingestion ensures seamless data flow into the Data Lakehouse

- [illegible] + Filebeat
- Logstash
- Kafka -
- Kafka Connect

On Premise? -> MSK [illegible]

02

Storage

The robust storage of the Data Lakehouse accommodates structured, unstructured, and semi-structured data, enabling scalable and efficient data management.

S3 + Parquet + Hive

Hive Metastore -> Glue

Iceberg

03

Processing

In the Data Lakehouse architecture, processing is efficiently [illegible] component. [illegible] we use Spark and Trino

[illegible]

Spark 3.5

Scalability and [illegible]

04

Orchestration

The numerous jobs we have, are coordinated/orchestrated by a component called [illegible]. [illegible] responsibility is to run the correct job at the correct time

More [illegible]

Hybrid [illegible]

Conclusion

To sum it up:
- Prezi Data Platform is founded on stable/state-of-the-art techs and architecture decision
- It allows the introduction of cutting edge technology, as it is pluggable and modular
- It's constantly evolving

Trino Summit 2024

Thanks!

@vincenzocassaro