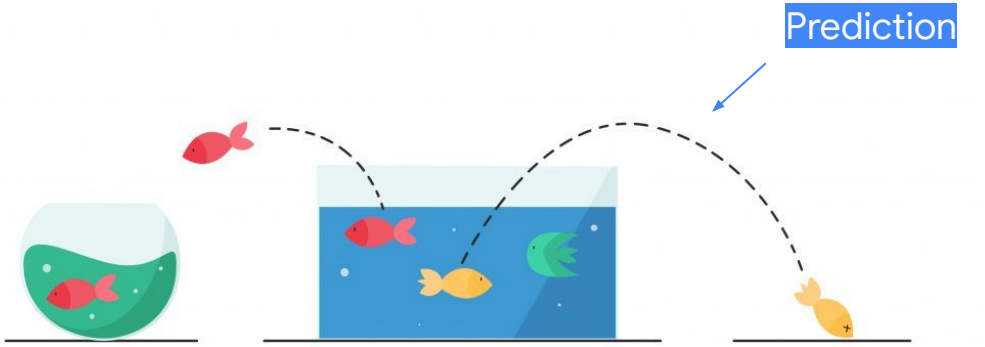
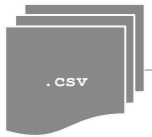
The job to be done:
A churn prediction model



The job to be done: Build a data product

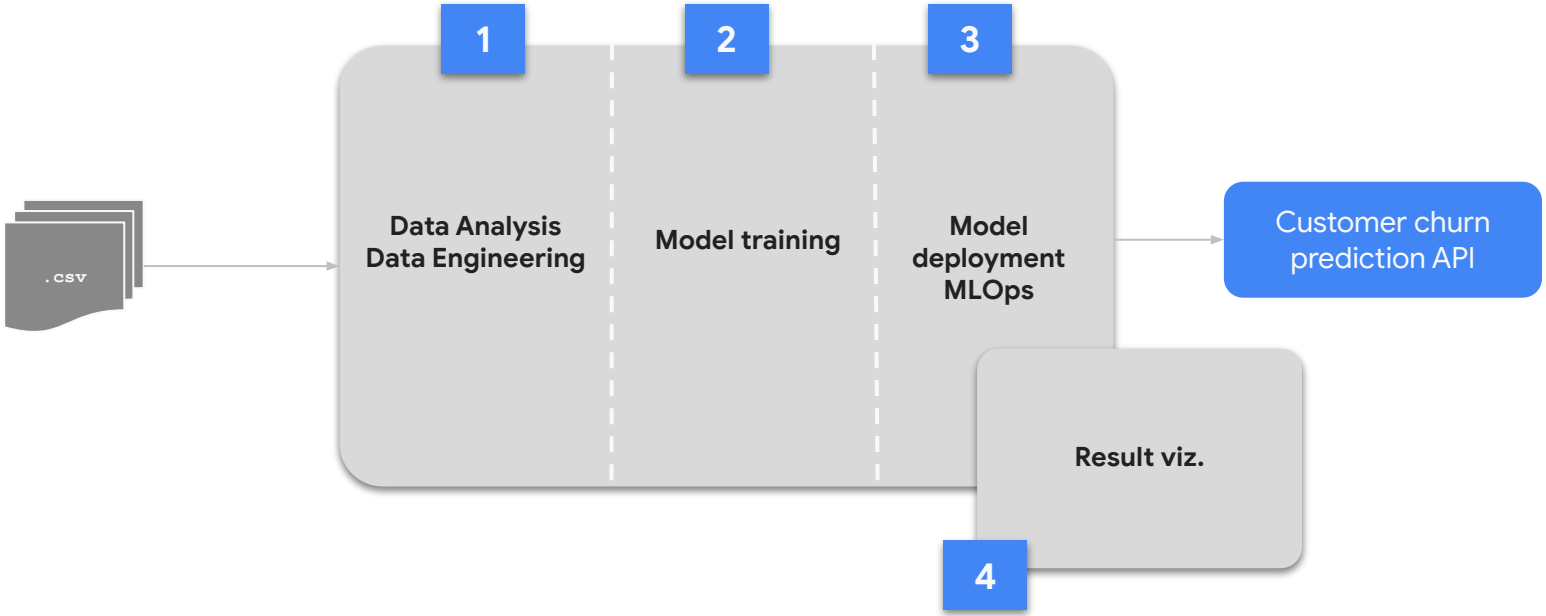
High value telco customer list variables and **churn tag (Y/N)**



Data system

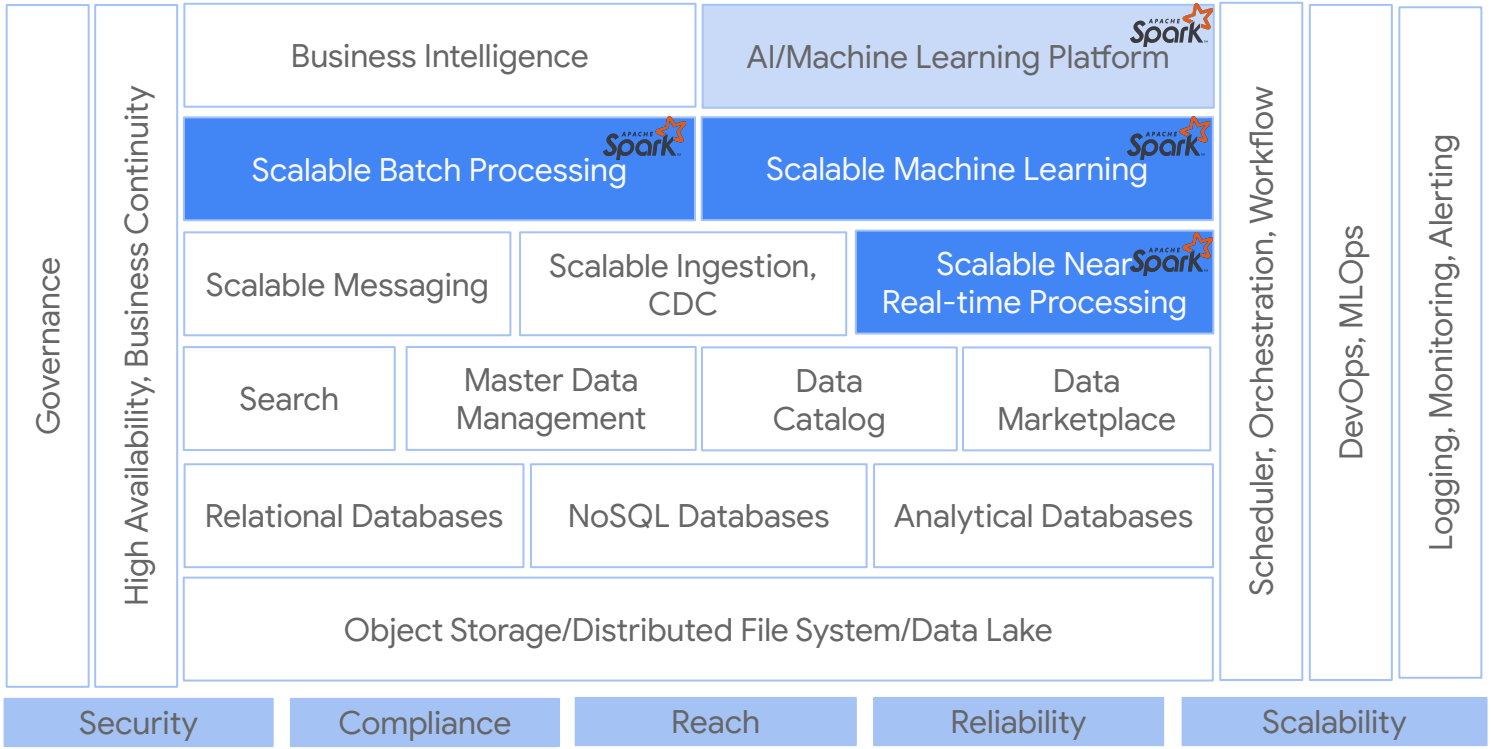
Customer churn prediction API

The job to be done: Product dev stages



The toolkit - SPARK

The toolkit


Governance
High Availability, Business Continuity
Business Intelligence
AI/Machine Learning Platform
Scalable Batch Processing
Scalable Machine Learning
Scalable Messaging
Scalable Ingestion, CDC
Scalable Near Real-time Processing
Search
Master Data Management
Data Catalog
Data Marketplace
Relational Databases
NoSQL Databases
Analytical Databases
Object Storage/Distributed File System/Data Lake
Scheduler, Orchestration, Workflow
DevOps, MLOps
Logging, Monitoring, Alerting
Security
Compliance
Reach
Reliability
Scalability

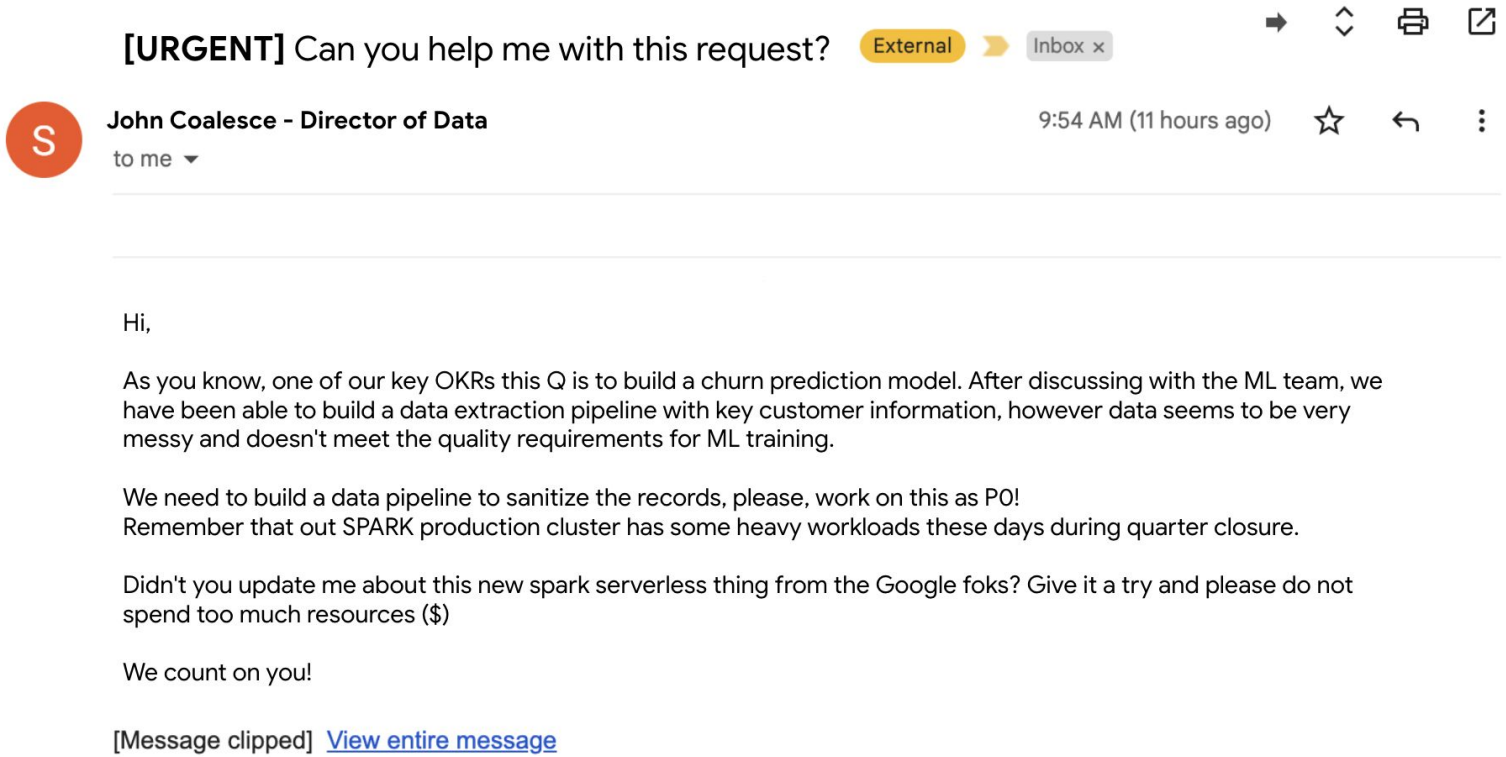
The personas

Data Analyst
Business Analyst

Data Engineer
ML Engineer

Data Scientist

The trigger - You've got email



[URGENT] Can you help me with this request? External Inbox ×

S **John Coalesce - Director of Data** 9:54 AM (11 hours ago)

to me

Hi,

As you know, one of our key OKRs this Q is to build a churn prediction model. After discussing with the ML team, we have been able to build a data extraction pipeline with key customer information, however data seems to be very messy and doesn't meet the quality requirements for ML training.

We need to build a data pipeline to sanitize the records, please, work on this as P0!
Remember that out SPARK production cluster has some heavy workloads these days during quarter closure.

Didn't you update me about this new spark serverless thing from the Google foks? Give it a try and please do not spend too much resources ($)

We count on you!

[Message clipped] View entire message

Where's the data?

Left it on a bucket
Good luck!

Telco customer list with **churn information**

1

2

3

Data engineering
Data preparation

Model training

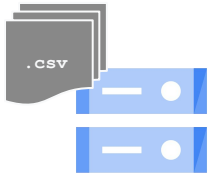
Model deployment
MLOps

Customer churn prediction API

Result viz.

4

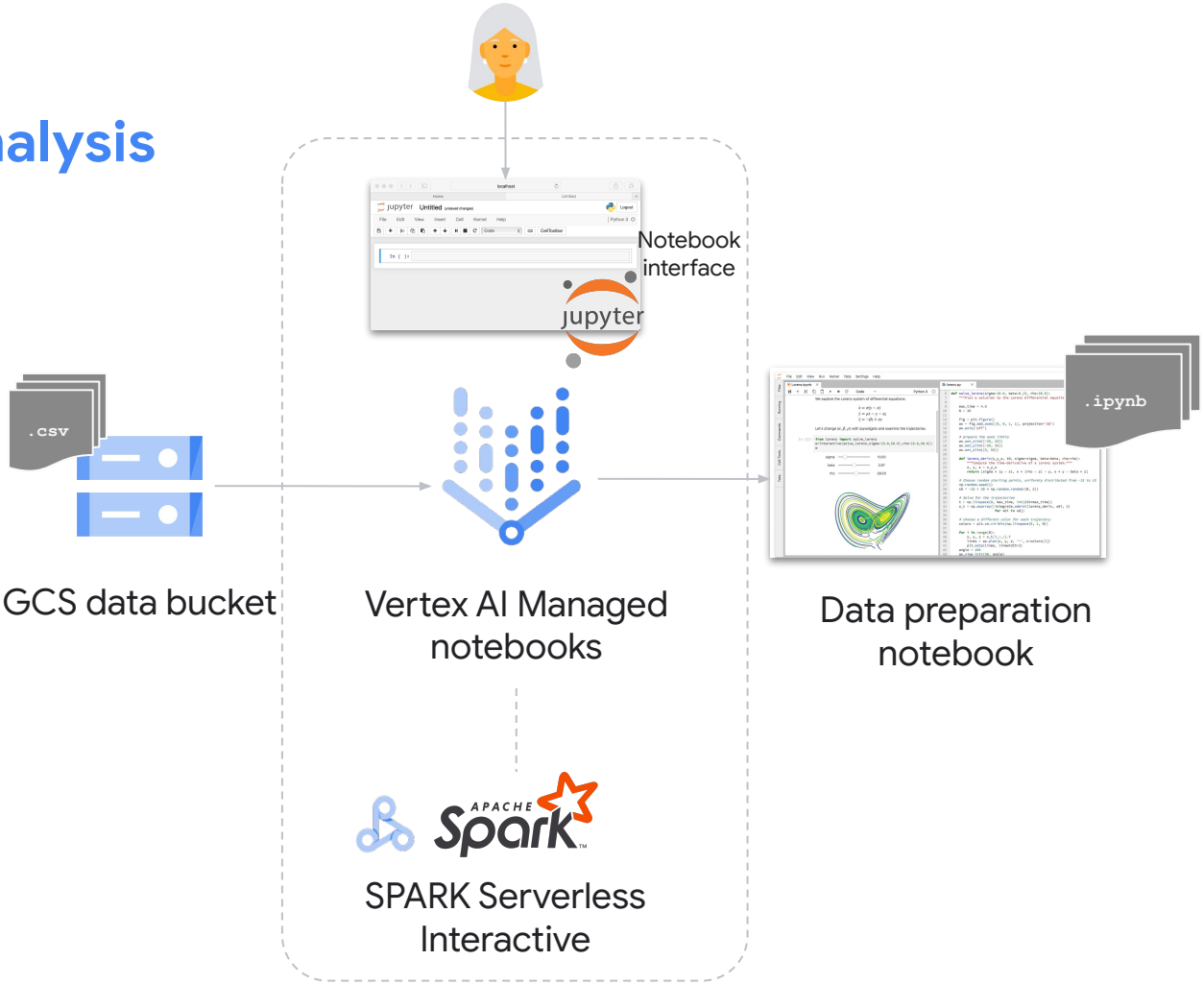
1 Data Analysis



GCS data bucket

- Time constrained
- Integrated with data sources
- Interactive data analysis (EDA)
- Able to data cleansing at scale
- No need to provision/prepare a large cluster
- Economic

1 Data Analysis



1 Data Engineering

Data preparation notebook

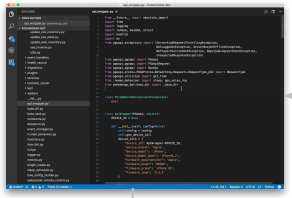
RAW data

- Repeatable process
- At scale
- Enterprise level: Monitored, secured, ..
- Portable
- No need to provision/prepare a large cluster
- Economic

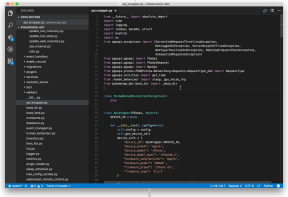
Prepared data

1 Data Engineering

IDE interface



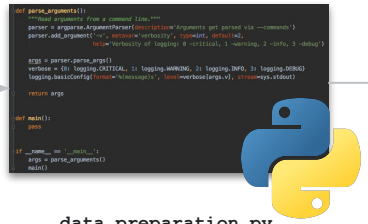
IDE interface



Data preparation notebook

Production Code refactor

Testing
Error handling
Argument Parsing
Performance
Security
Libraries
...



`data_preparation.py`

Pipeline code

RAW data

Composer
dataproc
serverless
operator

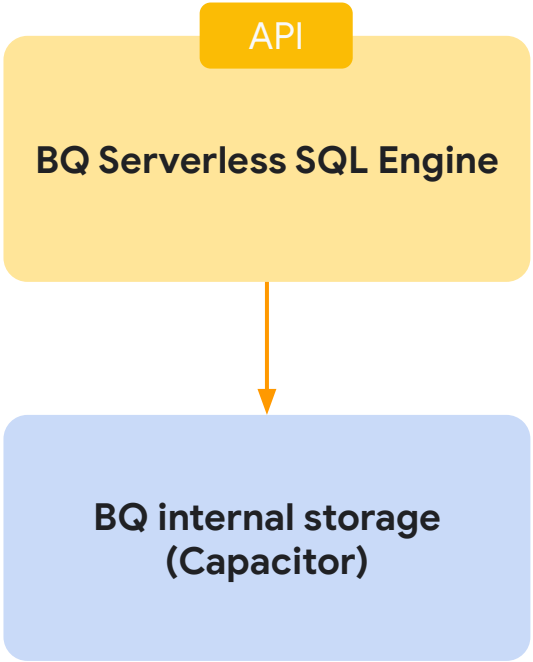
dbt python models

BigQuery Spark stored procedure

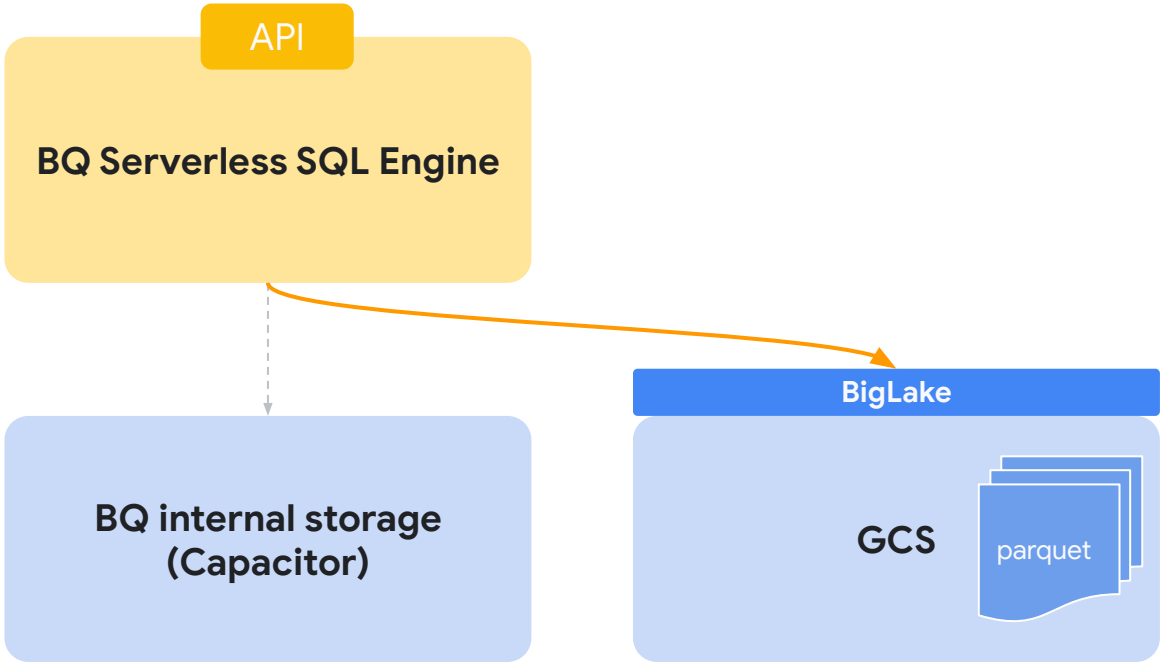
Serverless Batch

Prepared data

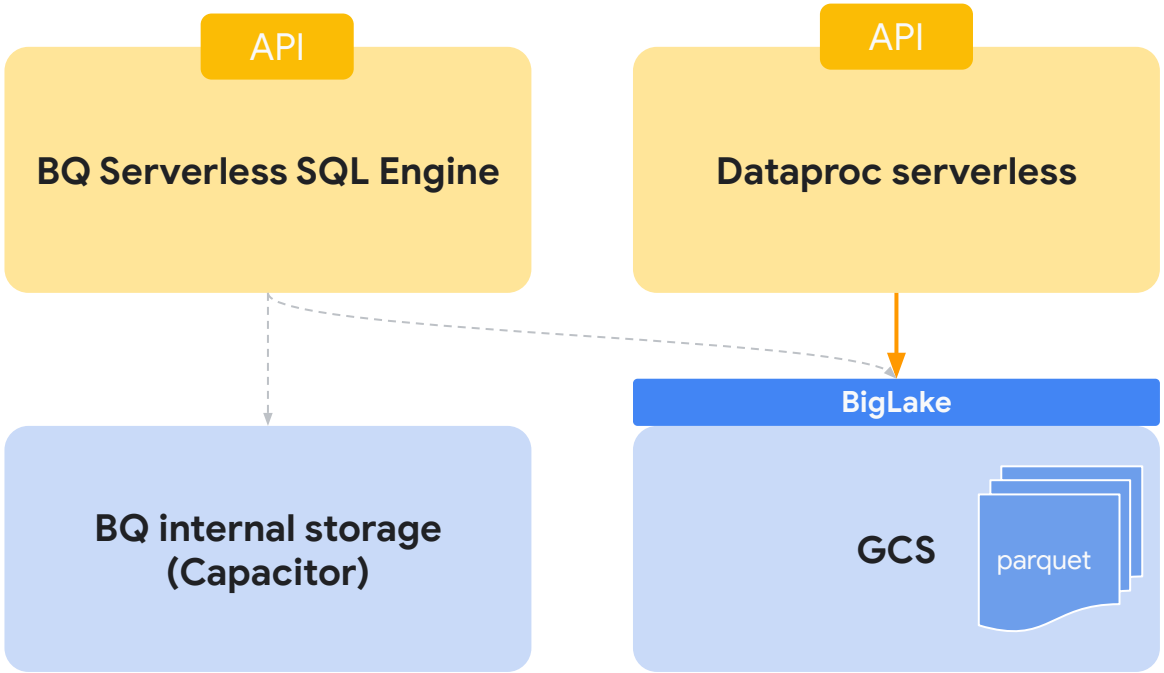
{Interlude} Interoperability



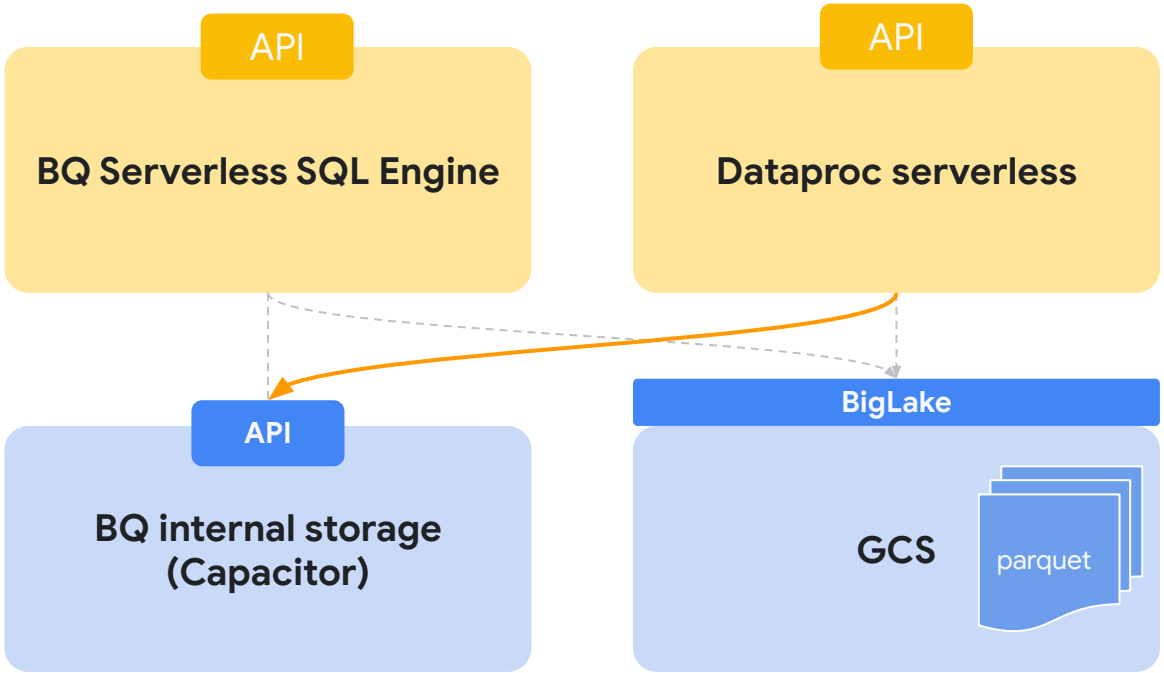
{Interlude} Interoperability



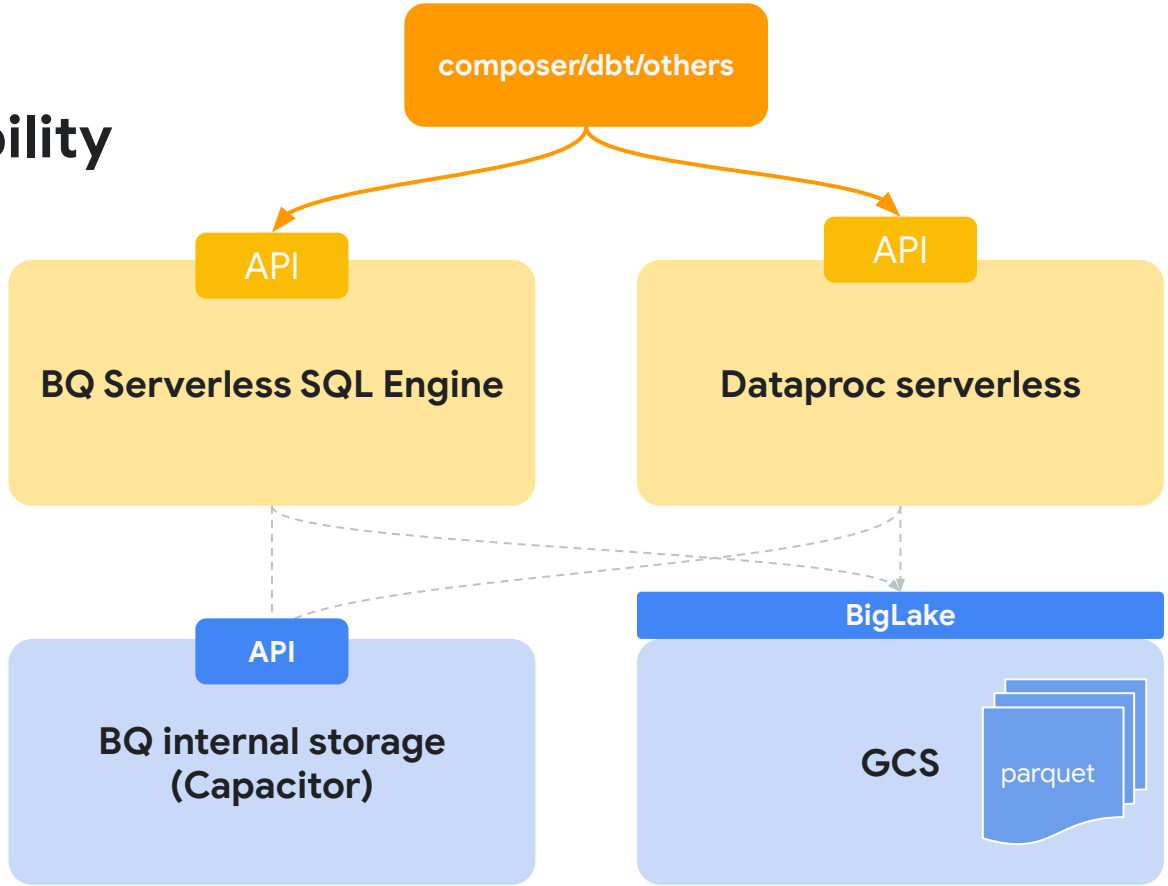
{Interlude} Interoperability



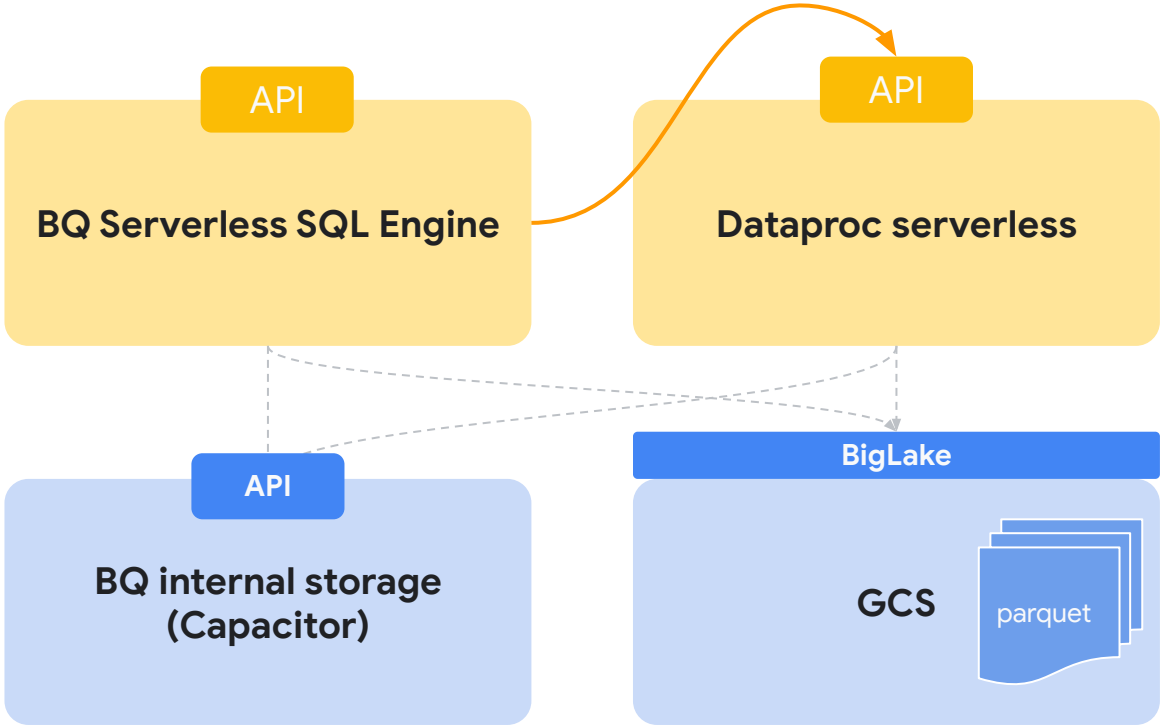
{Interlude} Interoperability



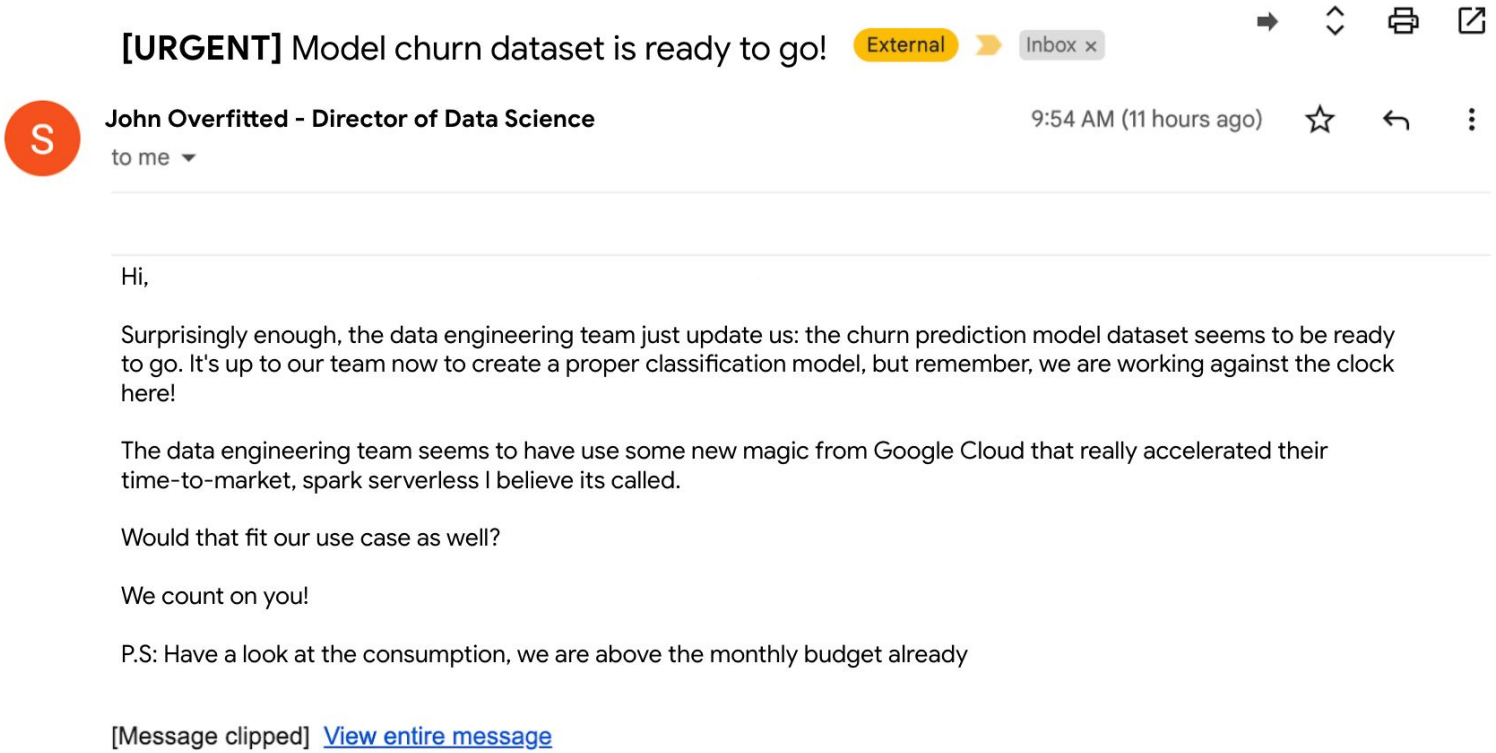
{Interlude}
Interoperability



{Interlude} Interoperability



The trigger



[URGENT] Model churn dataset is ready to go! External Inbox ×

John Overfitted - Director of Data Science 9:54 AM (11 hours ago)

to me

Hi,

Surprisingly enough, the data engineering team just update us: the churn prediction model dataset seems to be ready to go. It's up to our team now to create a proper classification model, but remember, we are working against the clock here!

The data engineering team seems to have use some new magic from Google Cloud that really accelerated their time-to-market, spark serverless I believe its called.

Would that fit our use case as well?

We count on you!

P.S: Have a look at the consumption, we are above the monthly budget already

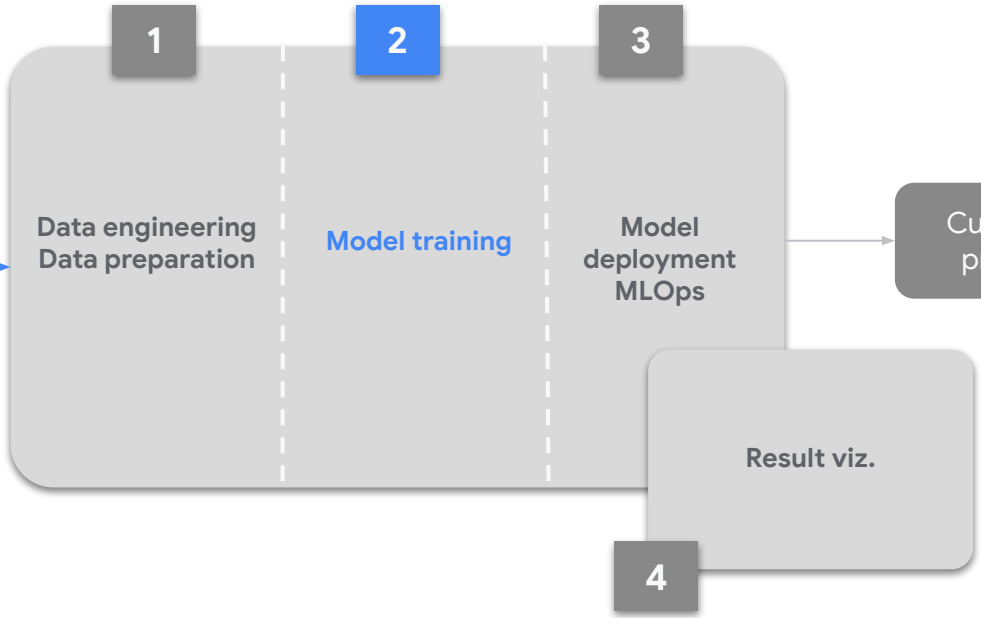
[Message clipped] View entire message

Whats next?

RAW data

Prepared data

Telco customer list with **churn information**



2 Model Building

BQ prepared data

- Integrated with data sources

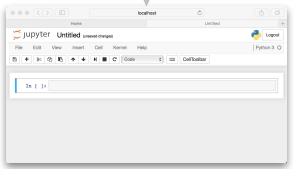
  Model training at scale

- No need to provision/prepare a large cluster

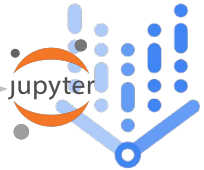
- Economic

2 Model Building

Notebook interface

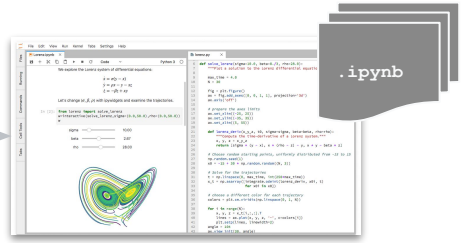


BQ prepared data



Vertex AI Managed notebooks

SPARK Serverless Interactive



Model training notebook

The trigger

[URGENT] Model churn to prod!

External Inbox ×

Maria DriftSkew - Director of ML Engineering 9:54 AM (11 hours ago)

to me

Hi,

Have your heard the news? The DS team has just released their churn prediction model! No delays this time!

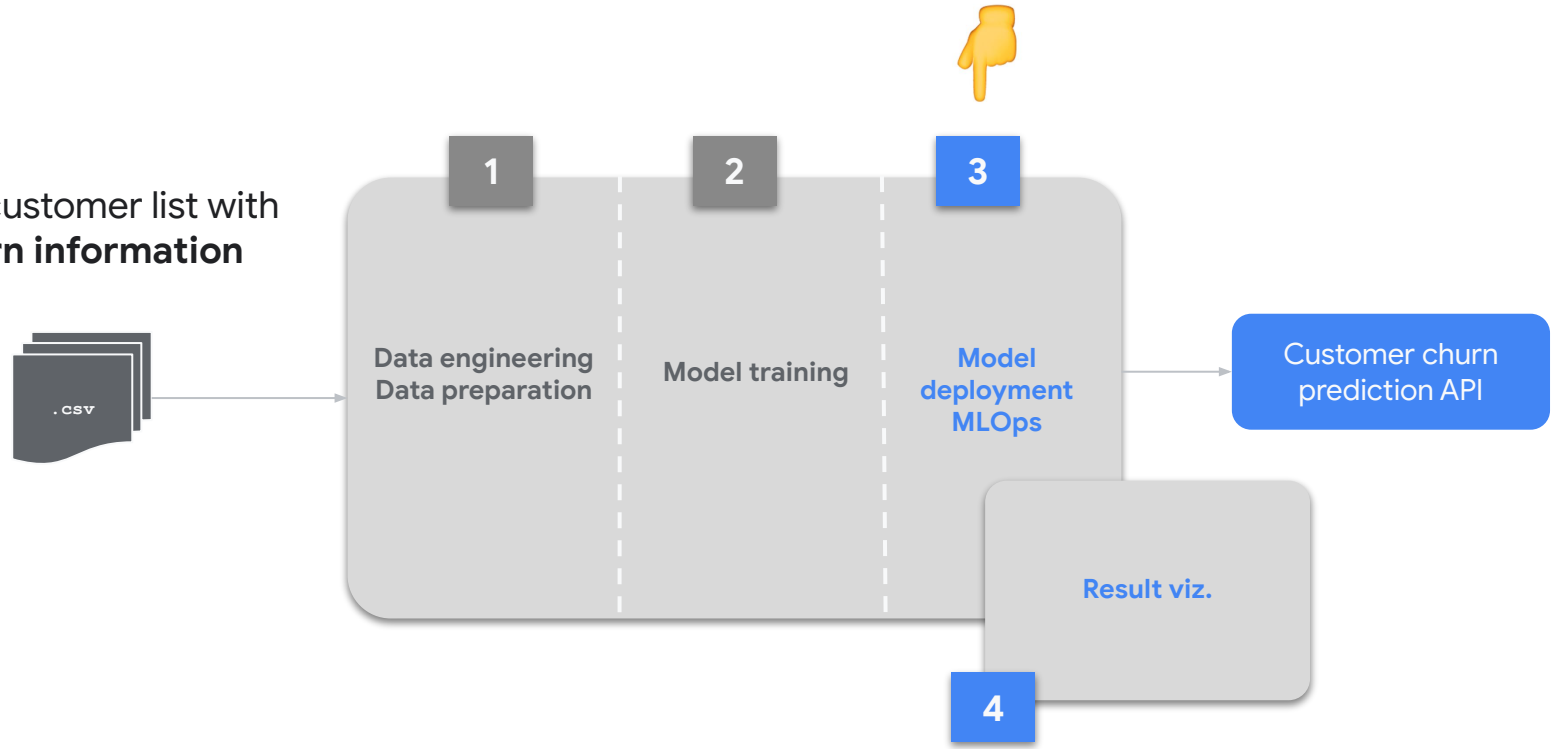
So, it's our turn now, we need to build the a ML pipeline to ensure the model is up to date!

Honestly, never deployed a SPARK MLlib model before, good luck!

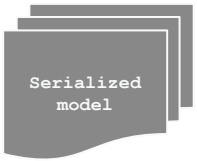
P.S: Feel free to use another pipeline tool other than Composer

[Message clipped] View entire message

Telco customer list with **churn information**



3 ML Ops



MLeap model

- Support for online inference

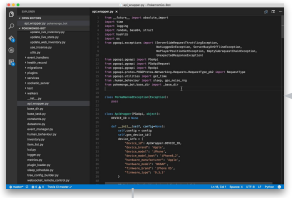
  Automated Model (re)-training at scale

- Rich metadata
- Serverless

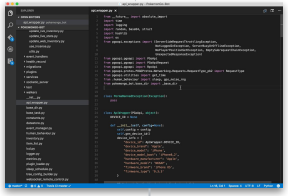
Customer churn prediction API

3 ML Ops

IDE interface



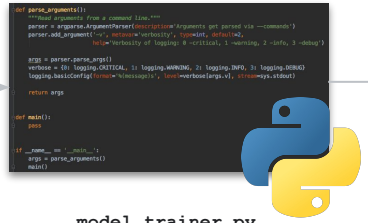
IDE interface



Model training notebook

Production Code refactor

Testing
Model serialization
Model deployment
Error handling
Argument Parsing
Performance
Security
Libraries
...



`model_trainer.py`

ML Pipeline code

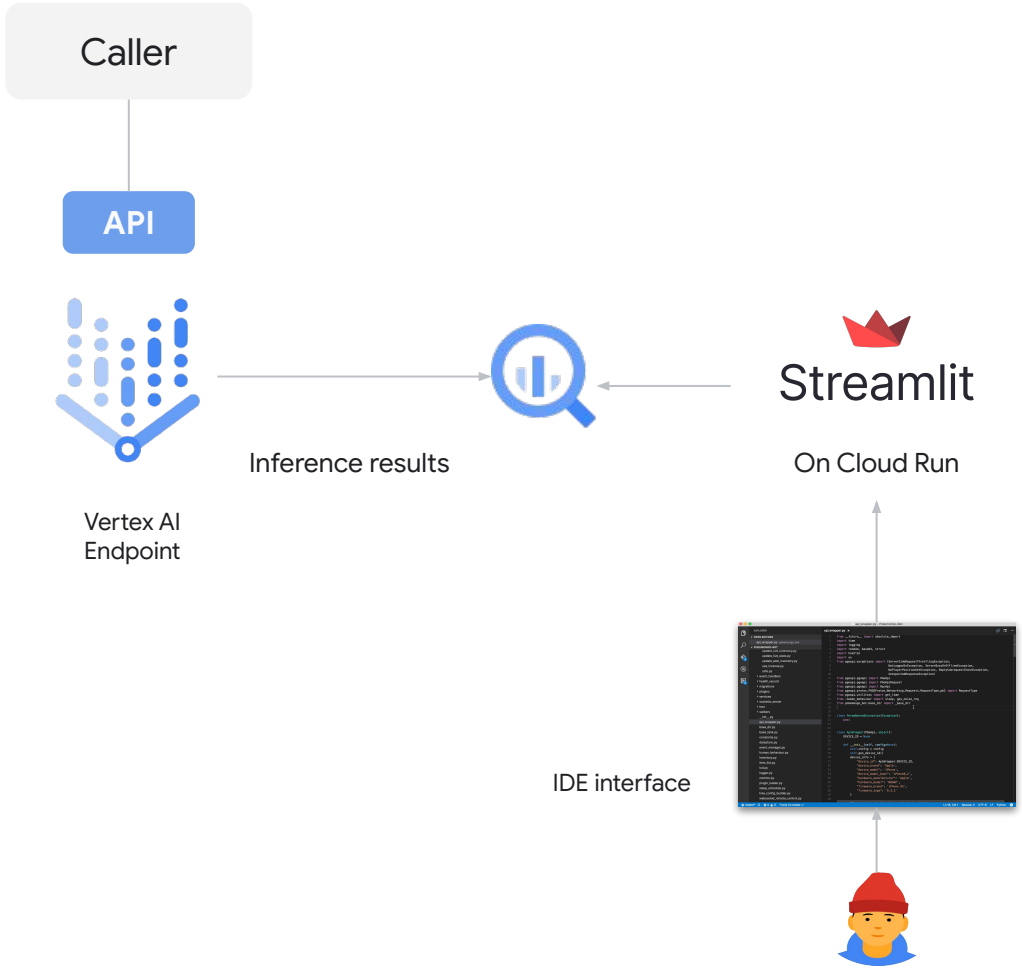
Vertex AI Pipelines

Prepared data

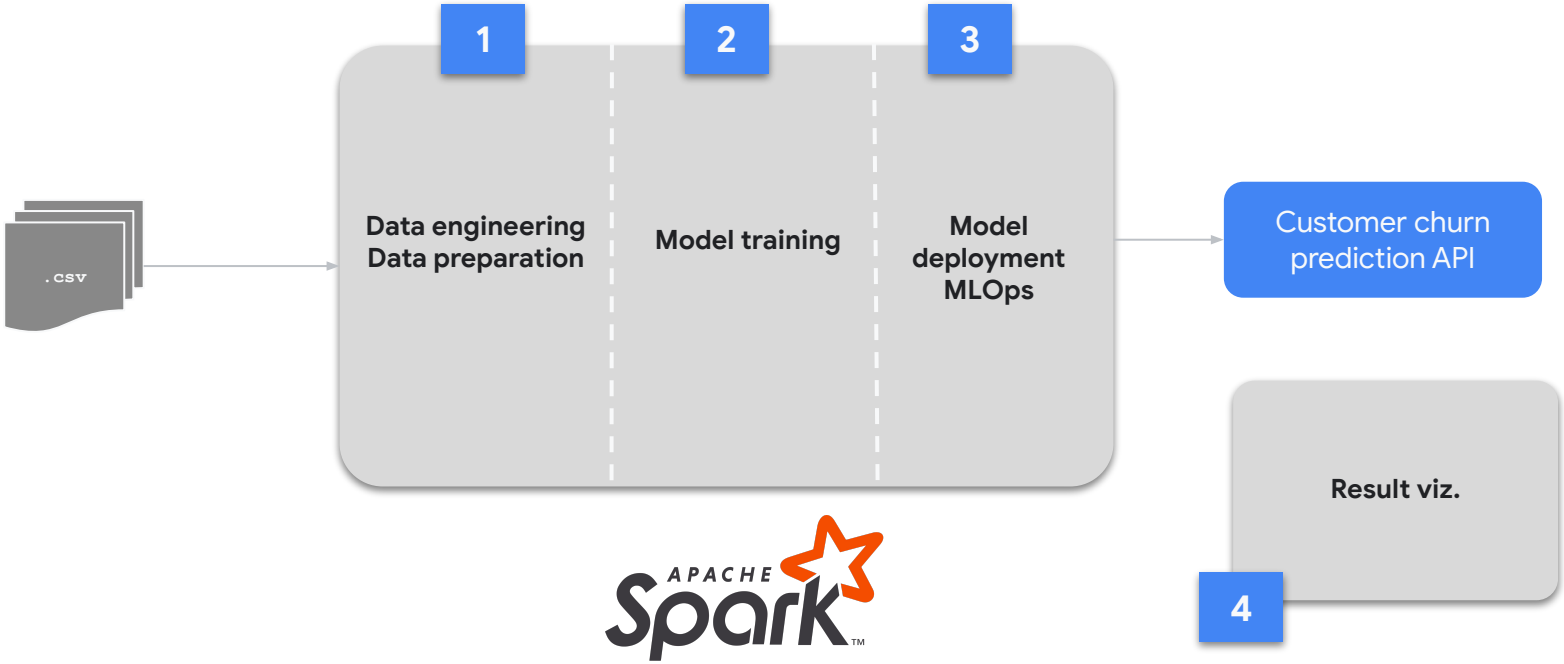
Serverless Batch

Model endpoint REST

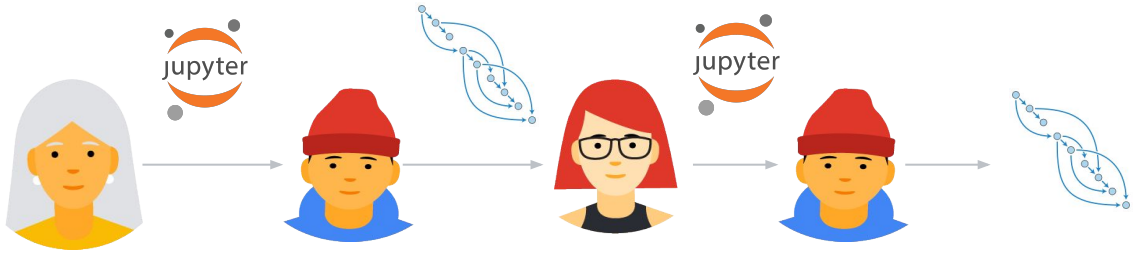
4 Data Viz

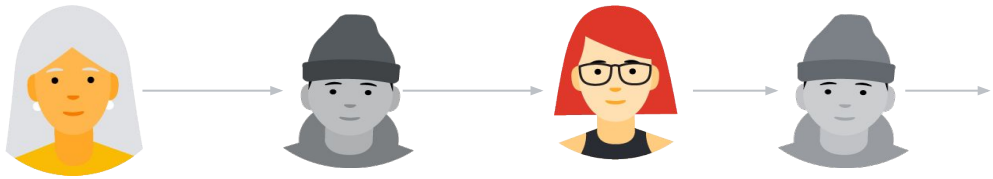

Caller
API
Inference results
Streamlit
On Cloud Run
Vertex AI
Endpoint
IDE interface

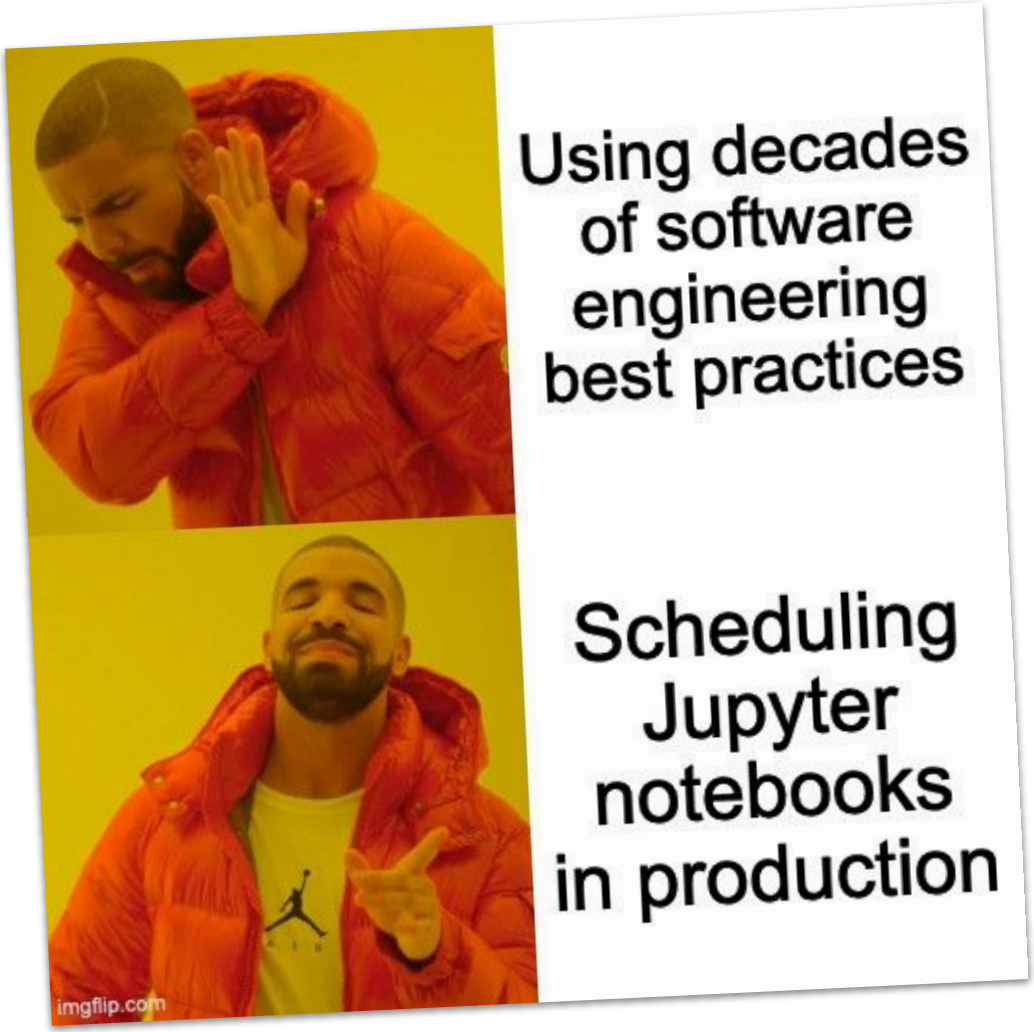
All in all


1
2
3
.csv
Data engineering
Data preparation
Model training
Model
deployment
MLOps
Customer churn
prediction API
Result viz.
4
APACHE
Spark

SPARK - A first class citizen in Google Cloud


jupyter
jupyter




Using decades of software engineering best practices
Scheduling Jupyter notebooks in production
imgflip.com