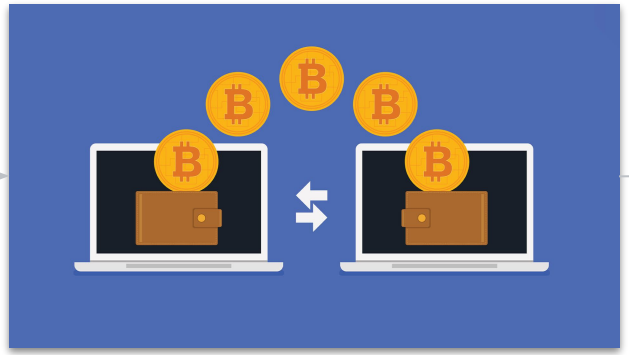
The job to be done: A mining pool classifier model

Sender



Receiver

Verification
(hash calculation)

Transaction “mined” by pool?

The job to be done: A mining pool classifier model

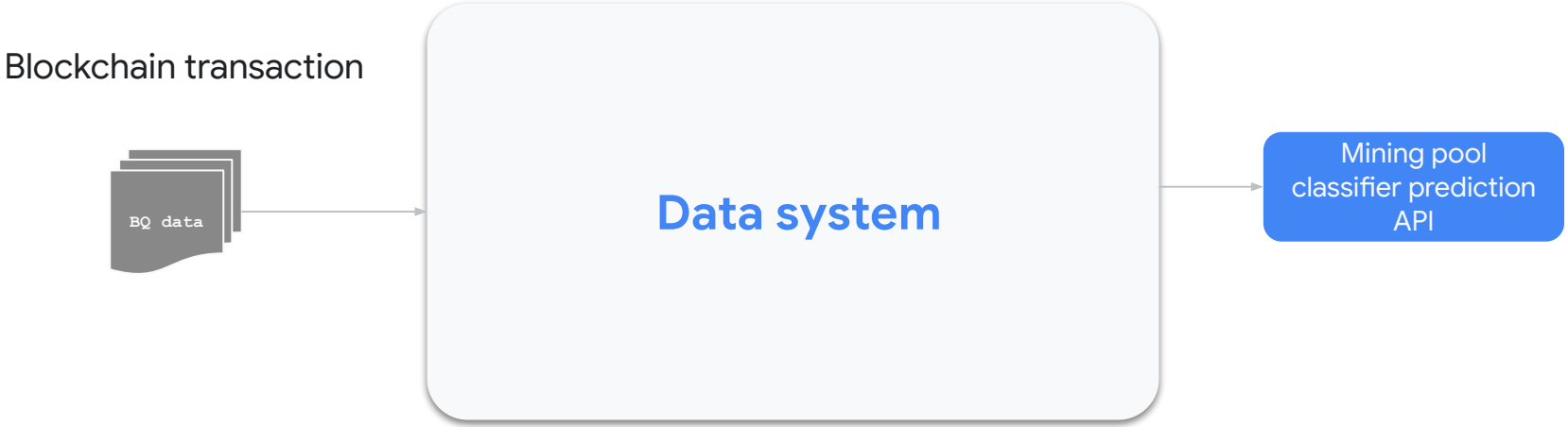
Address classification

Blockchain transaction history can be aggregated by address and used to analyze user behavior. To motivate further exploration, we present a simple classifier that can detect Bitcoin mining pools. As a brief historical note, mining pools were created when the difficulty of mining Bitcoin reached such a level that rewards could be expected only once every few years. Miners began to pool their resources to earn a smaller share of rewards more consistently and in proportion to their contribution to the pool in which they were mining.

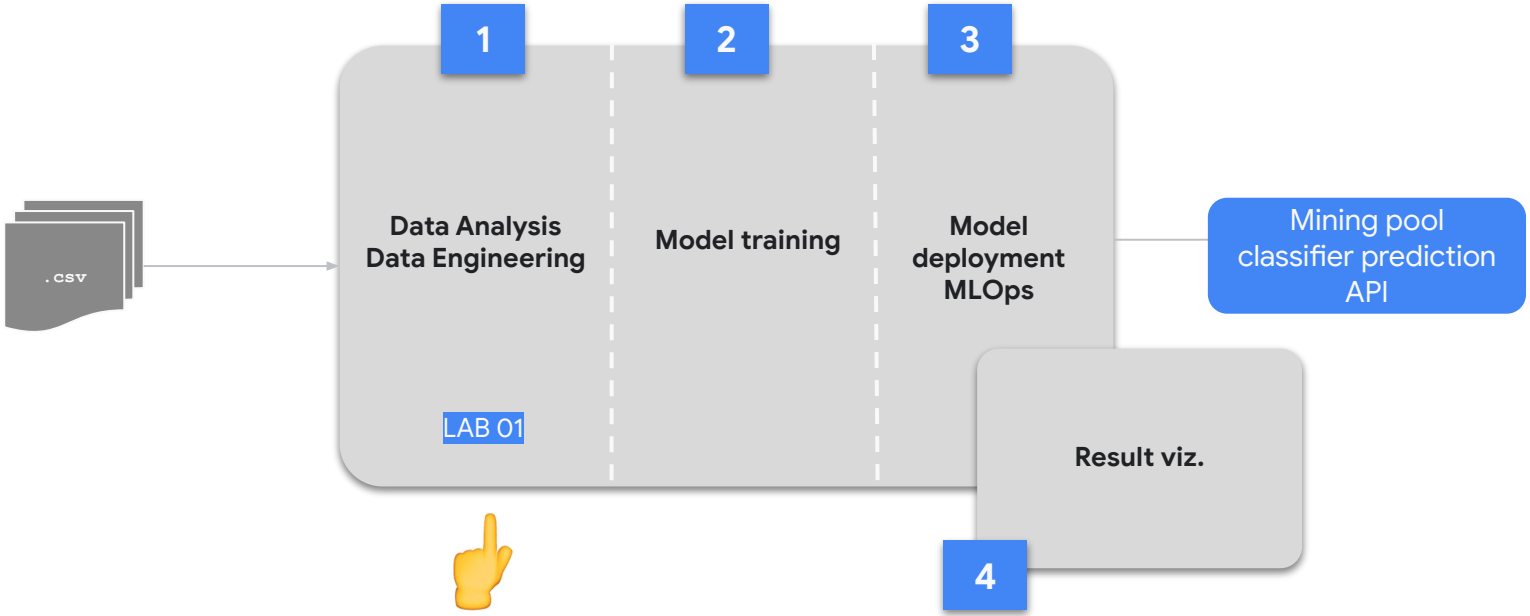
First, we constructed 26 feature vectors to characterize incoming and outgoing transaction flows to each address. Next, we trained the model using labels derived from transaction signatures. Many large mining pools identify themselves in the signature of blocks' Coinbase transactions. Parsing these signatures, we labelled 10,000 addresses as belonging to known mining pools. One million other addresses were included in the dataset as "non-miners." The query used to generate our features and labels can be seen here, and the source code for this analysis can be found in a Kaggle notebook here.

BQ crypto currencies public datasets

The job to be done: Build a data product

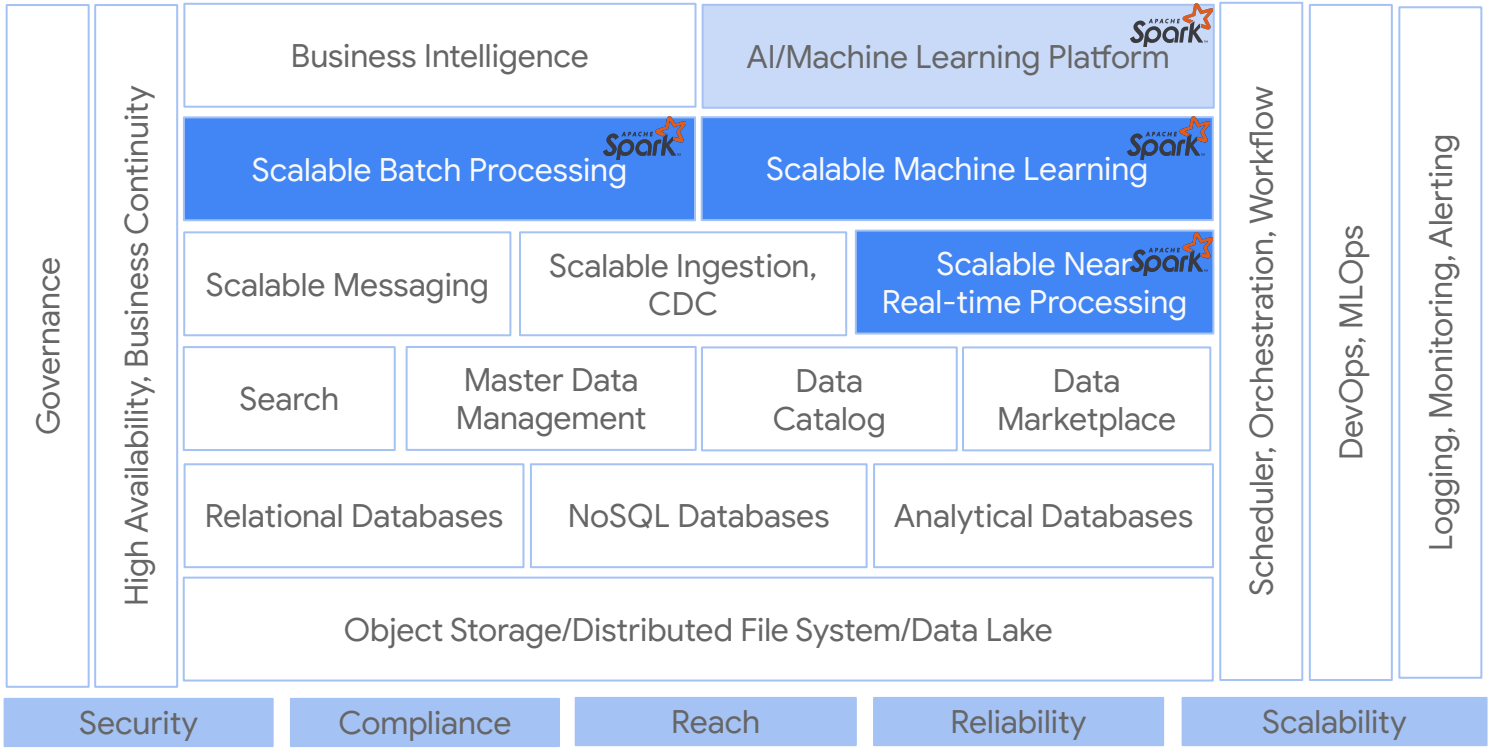


The job to be done: Product dev stages



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

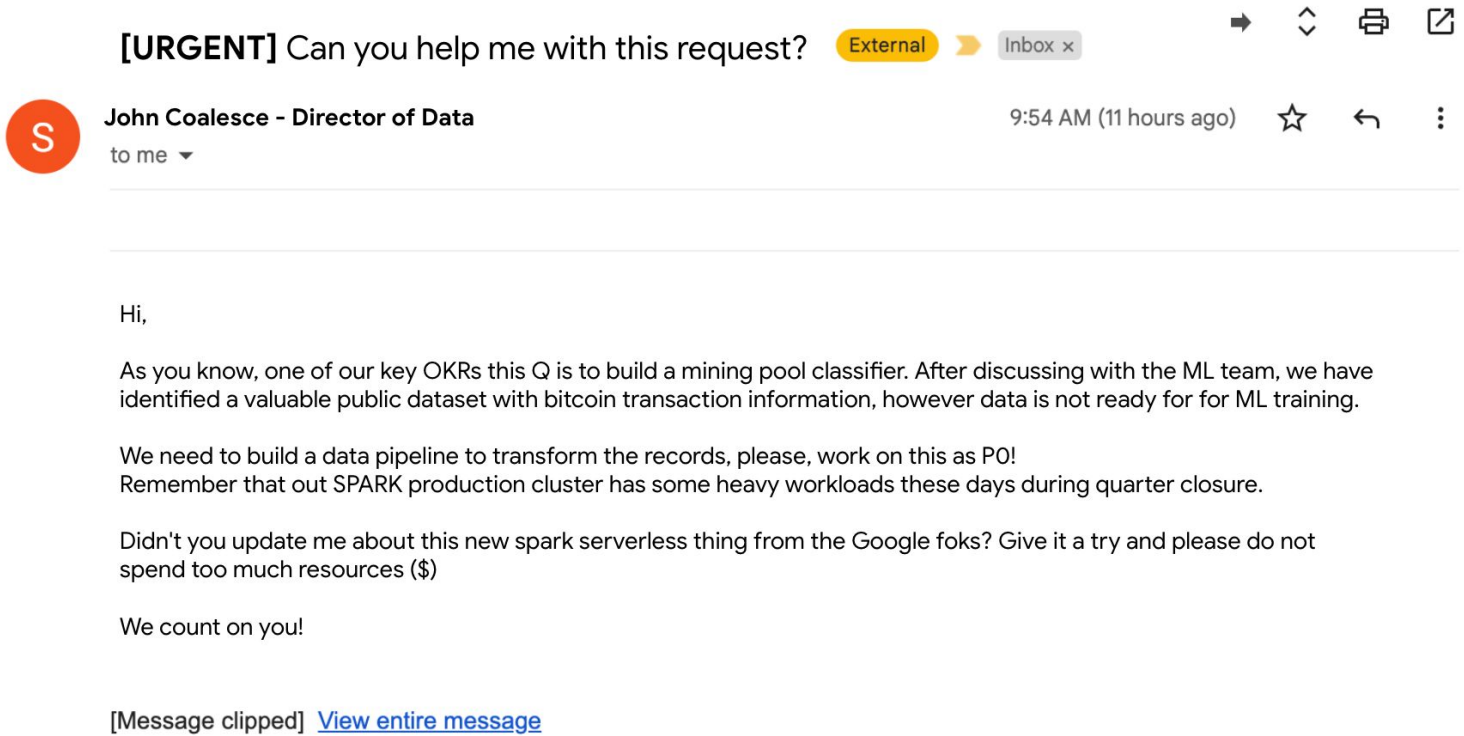
The personas

Data Analyst
Business Analyst

Data Engineer
ML Engineer

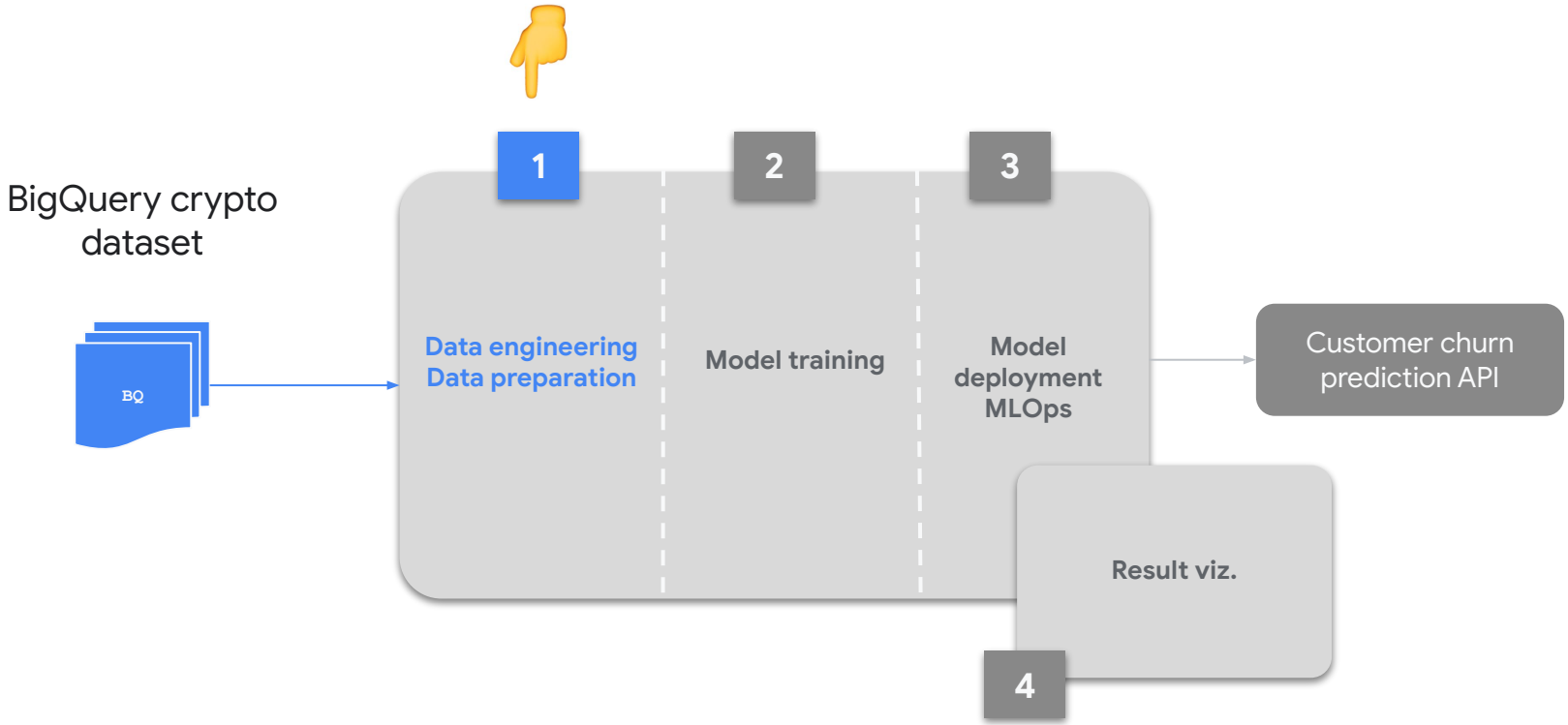
Data Scientist

The trigger - You've got email



Where's the data?

Check on BigQuery
Good luck!

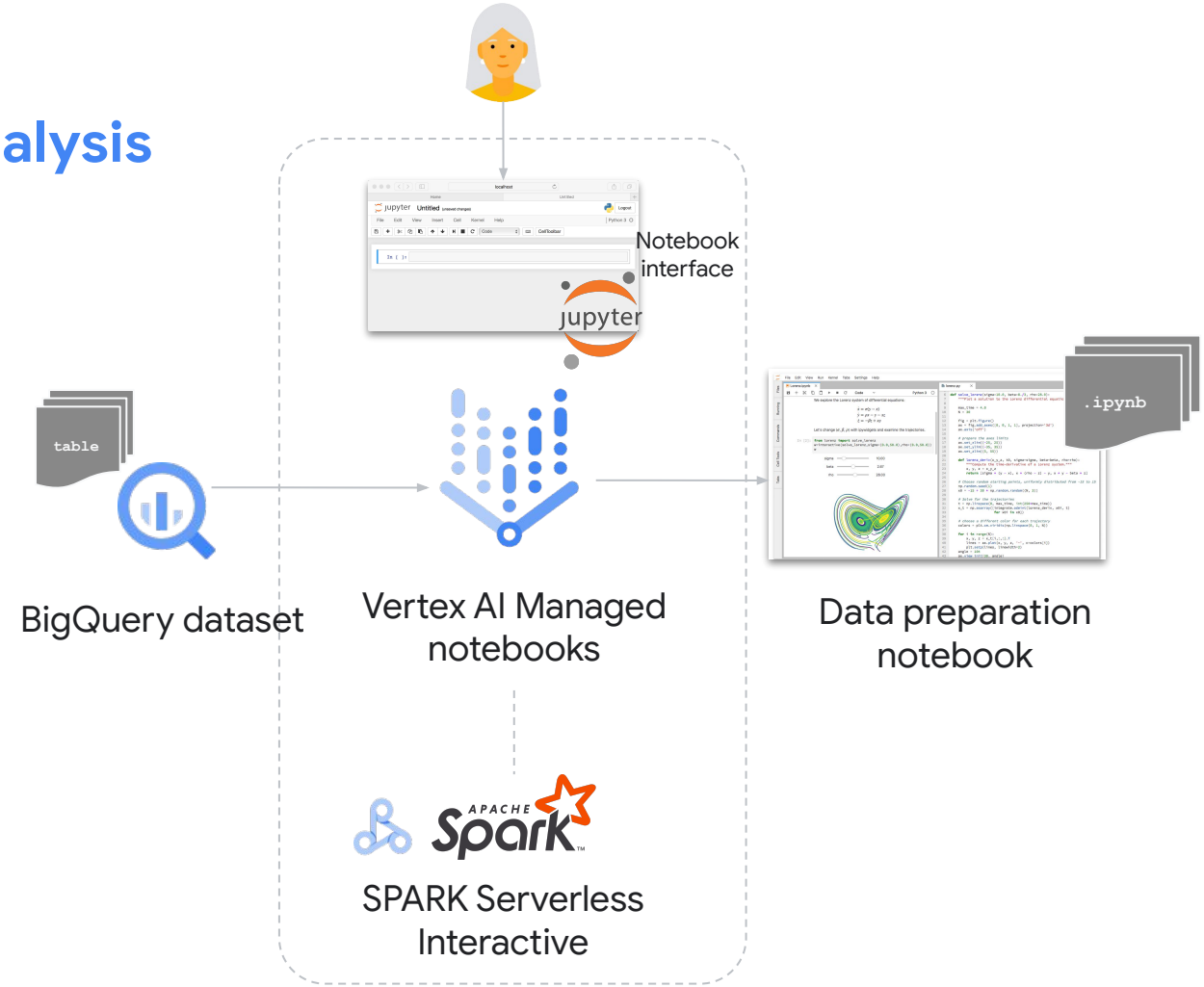

BigQuery crypto dataset
BQ
1
2
3
Data engineering
Data preparation
Model training
Model deployment
MLOps
Customer churn prediction API
Result viz.
4

1 Data Analysis

BigQuery dataset

- 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

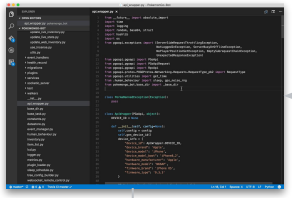
RAW data

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

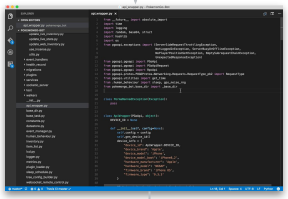
Prepared data

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

Composer
dataproc
serverless
operator

RAW data

Serverless Batch

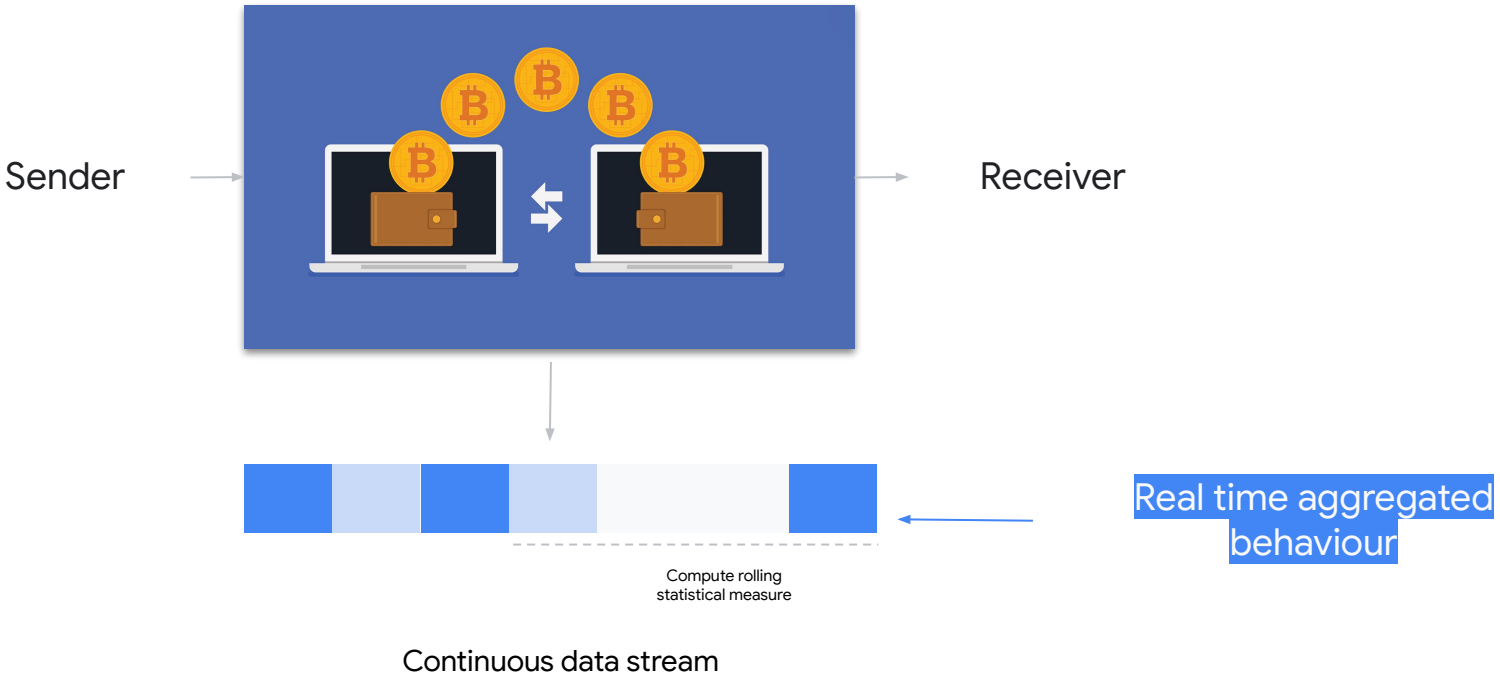
Prepared data

Whats next?

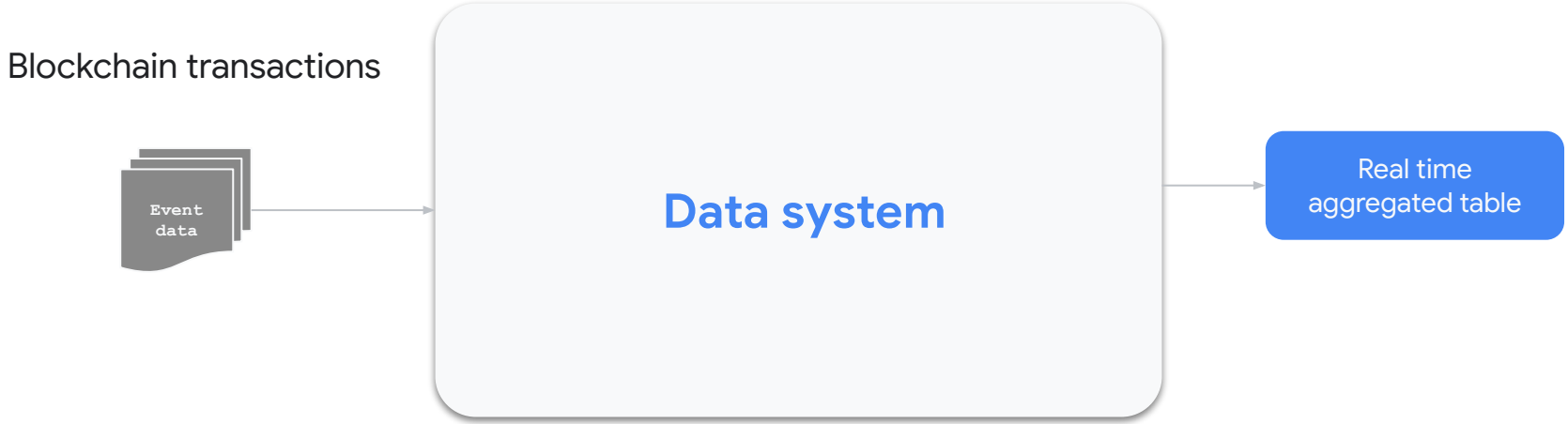
BigQuery dataset

Prepared data

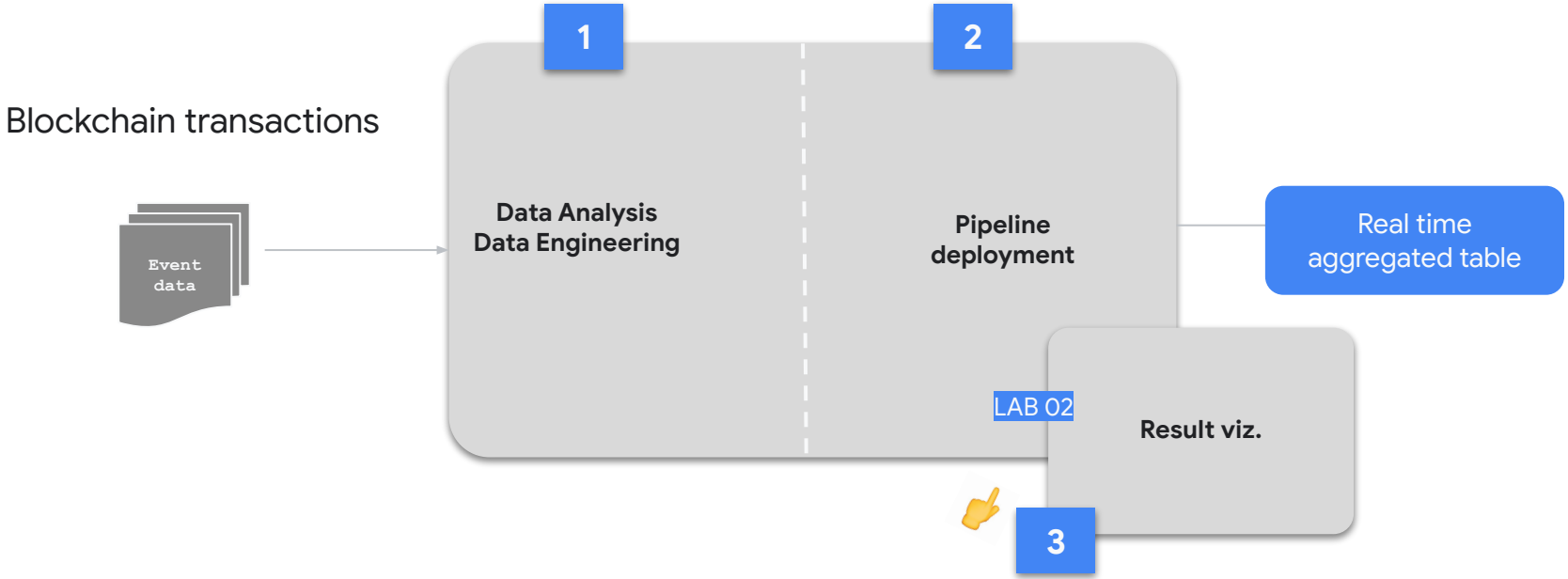
The job to be done: A real time analytics dashboard



The job to be done: Build a data product

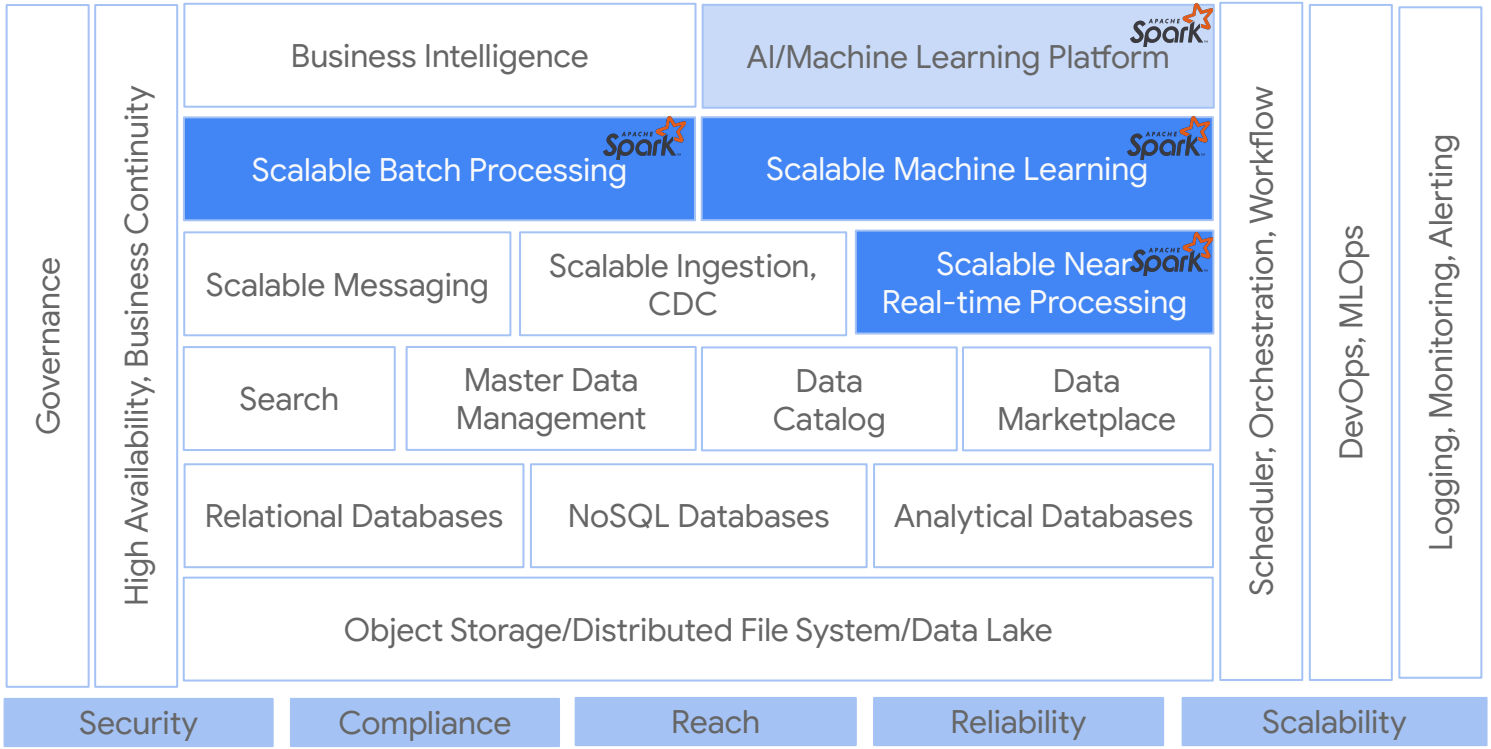


The job to be done: Product dev stages



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

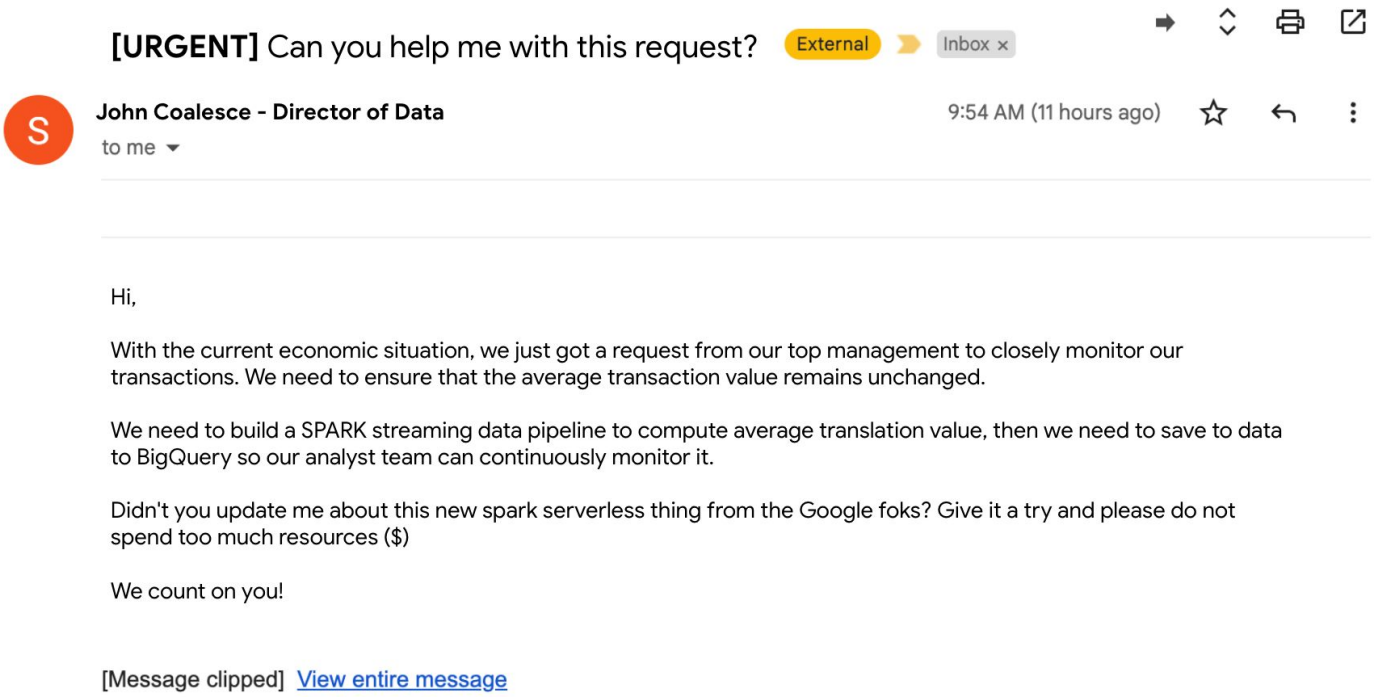
The personas

Data Analyst
Business Analyst

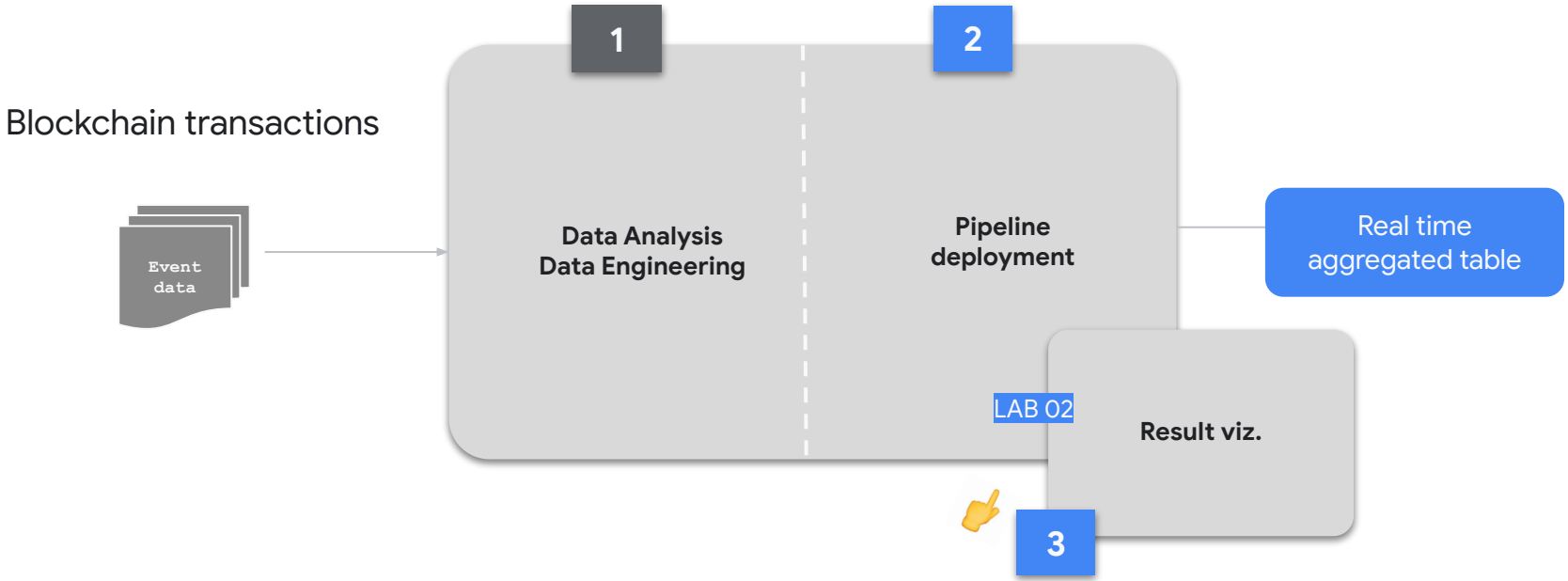
Data Engineer
ML Engineer

Data Scientist

The trigger - You've got email



The job to be done: Product dev stages



1 Pipeline deployment

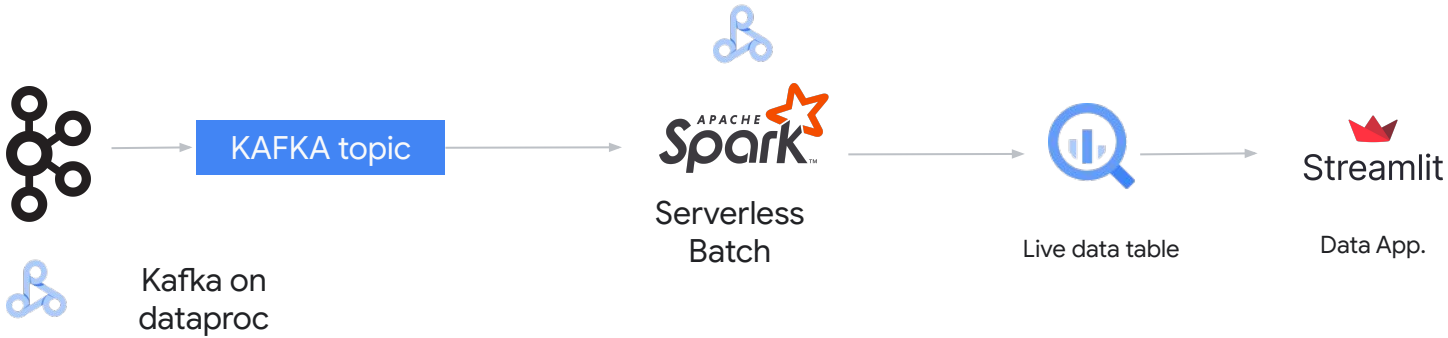
Ad hoc queries

RAW data

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

Live data table

1 Pipeline deployment



SPARK - A first class citizen in Google Cloud