Data Science with BigQuery

by Jonatan Reiners

The Agenda

- Our Problem
- Data Science
- BigQuery
- Big Picture / Data Pipeline
- Collecting Data
- Processing Data

Our Problem

- Analysis of App Market
- Various data sources
- 1.3 Million Apps * 58 Countries * Daily Data
- Volatile Information / Constant Improvement

Data Science

- Linear Regression
- Clustering, Grouping
- Modeling in R and Python
- Estimation in BigQuery

Why BigQuery?

- Structured Data
- Cheap Storage
- Columnar Storage is fast
- No Maintenance
- Wide Distribution

BigQuery

- REST API
- Things can fail ALWAYS!
- It’s NOT SQL! It’s DREMEL.
- From MAP REDUCE to DREMEL

Data Pipeline

- Import with Files
  - JSON nested Data
  - Cloud Storage
  - Backup

- BigQuery
    - RAW input tables
    - Processed Input
    - Estimation
    - Aggregation

Tools

- **pppusher**
  - import files in a structured Way
- **Bigrunner**
  - run processes/jobs in BigQuery
  - parallel Queries / scheduled Jobs
- **Data team tool belt**

Process input data

- JOIN EACH
  - Nested not chained
- be explicit with naming
- tables with day suffix and field

Thank You!

- It was a pleasure to present you the awesomeness of BigQuery
- please contact me if you have any questions
- jonatan@prioridata.com