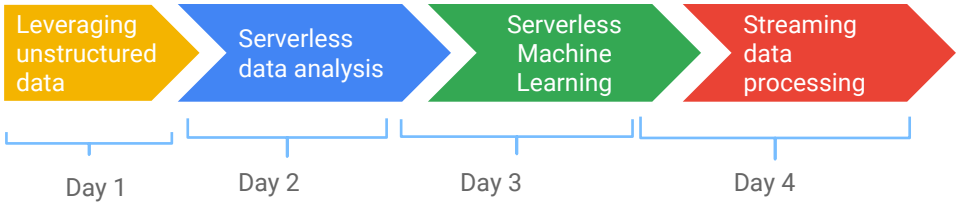
Google Cloud

Summary

Data Engineering on Google Cloud Platform

30 minutes

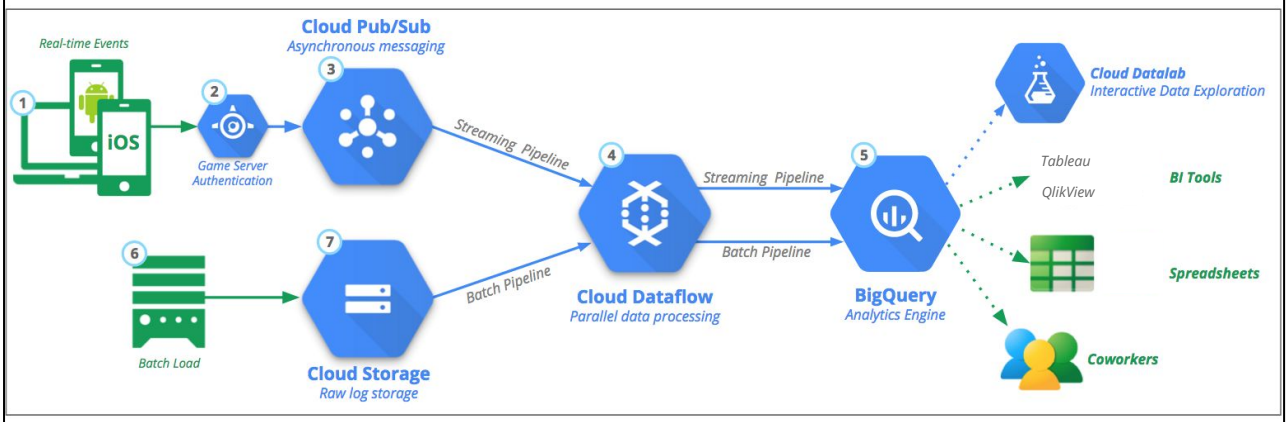
We covered a lot of ground...



Notes:

The ML content is approximately 1.5 days from end of Day 2 to start of Day 4. Timelines are approximate, of course.

Example architecture for data analytics



[Build a mobile gaming analytics platform - a reference architecture]

Notes:

Popular mobile games can attract millions of players and generate terabytes of game-related data in a short burst of time. This places extraordinary pressure on the infrastructure powering these games and requires scalable data analytics services to provide timely, actionable insights in a cost-effective way.

To address these needs, a growing number of successful gaming companies use Google's web-scale analytics services to create personalized experiences for their players. They use telemetry and smart instrumentation to gain insight into how players engage with the game and to answer questions like: At what game level are players stuck? What virtual goods did they buy? And what's the best way to tailor the game to appeal to both casual and hardcore players?

A new reference architecture describes how you can collect, archive and analyze vast amounts of gaming telemetry data using Google Cloud Platform's data analytics products. The architecture demonstrates two patterns for analyzing mobile game events:

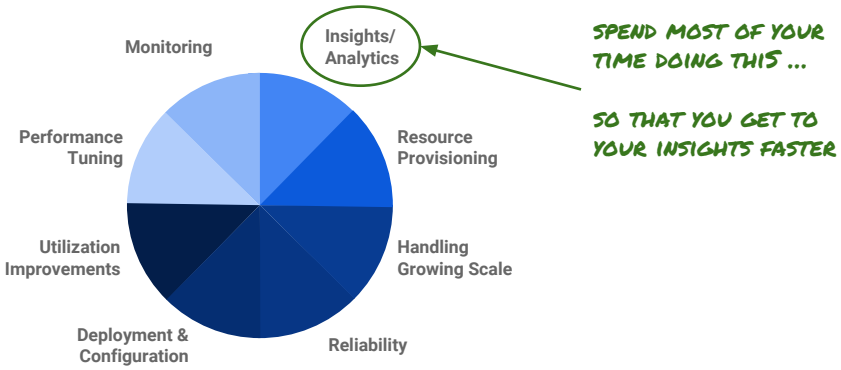
- Batch processing: This pattern helps you process game logs and other large files in a fast, parallelized manner. For example, leading mobile

- gaming company DeNA moved to BigQuery from Hadoop to get faster query responses for their log file analytics pipeline.
- Real-time processing: Use this pattern when you want to understand what's happening in the game right now. Cloud Pub/Sub and Cloud Dataflow provide a fully managed way to perform a number of data-processing tasks like data cleansing and fraud detection in real-time. For example, you can highlight a player with maximum hit-points outside the valid range. Real-time processing is also a great way to continuously update dashboards of key game metrics, like how many active users are currently logged in or which in-game items are most popular.

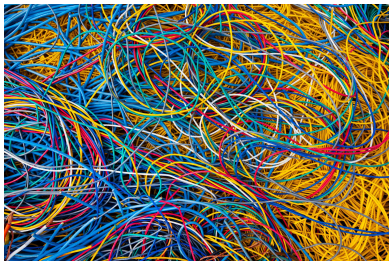
Some Cloud Dataflow features are especially useful in a mobile context since messages may be delayed from the source due to mobile Internet connection issues or batteries running out. Cloud Dataflow's built-in session windowing functionality and triggers aggregate events based on the actual time they occurred (event time) as opposed to the time they're processed so that you can still group events together by user session even if there's a delay from the source.

But why choose between one or the other pattern? A key benefit of this architecture is that you can write your data pipeline processing once and execute it in either batch or streaming mode without modifying your codebase. So if you start processing your logs in batch mode, you can easily move to real-time processing in the future. This is an advantage of the high-level Cloud Dataflow model that was released as open source by Google.
Cloud Dataflow loads the processed data into one or more BigQuery tables. BigQuery is built for very large scale, and allows you to run aggregation queries against petabyte-scale datasets with fast response times. This is great for interactive analysis and data exploration, like the example screenshot above, where a simple BigQuery SQL query dynamically creates a Daily Active Users (DAU) graph using Google Cloud Datalab.

Dataproc helps you focus on insights & analytics



Use ML APIs to analyze unstructured data



Images
Audio
Video
Free-form text

ML API

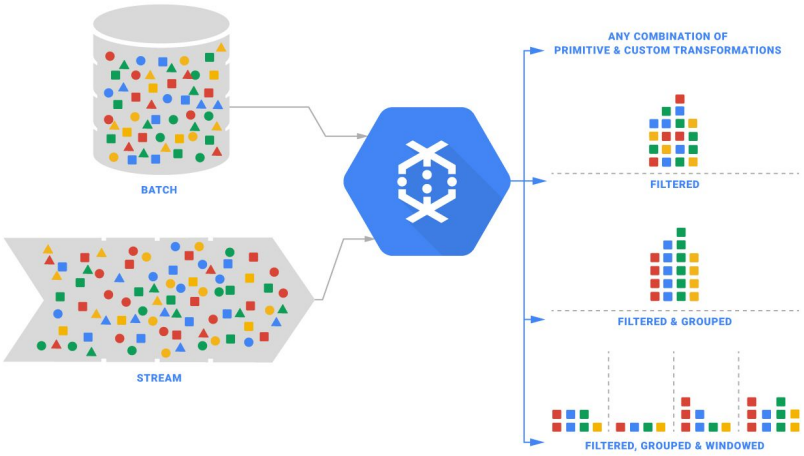
Places
Labels
People
Events
...

Cloud ML Engine

Notes:

https://pixabay.com/en/list-zettelbox-note-leaves-stack-1925395/ (cc0)

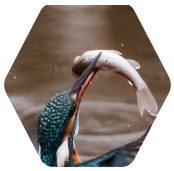
Dataflow does ingest, transform, and load



Notes:

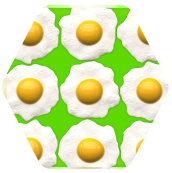
You can replace all the various data handling tools with just Dataflow. Many Hadoop workloads can be done easily and more maintainably with Dataflow. Plus, Dataflow is NoOps.

BigQuery provides interactive analysis



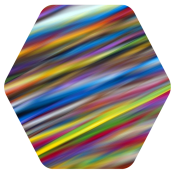
Near-real time analysis of massive datasets

No-ops;
Pay for use



Durable (replicated), inexpensive storage

Immutable audit logs



Mashing up different datasets to derive insights

Notes:

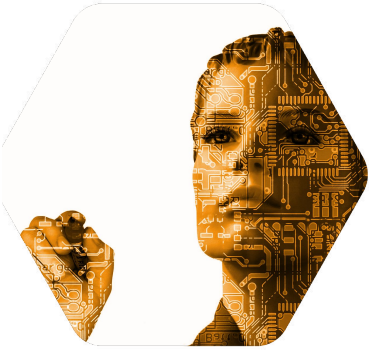
BigQuery provides near-real time queries. If you use Hadoop-like systems, you may have to wait an entire weekend for something that you can do in real-time in BigQuery, and this can allow your business to be far more agile and nimble. Think about changing prices, ordering more supplies or buying options every hour in response to demand rather than once a week. BigQuery is one of the transformative Google technologies -- it's no-ops. There is no cluster or software to maintain. You submit a query and you pay for the compute nodes only for the duration of that query. You don't have to pay to keep a compute cluster up and running. Whenever you ingest data into BigQuery, it's auto-replicated, and you get that reliability built-in to the cost of the storage. It's on the order of Nearline storage (i.e. less than standard cloud storage costs!). Google Cloud Audit Logs track Admin Activity and Data Access. These Immutable logs can tell you "who did what, where, and when?" in BigQuery.

Finally, one of the really transformational things is that because it is completely no-ops and a common query format, you can use BigQuery as the way to collaborate more within your company, to tie together datasets from across sales, catalog, warehouse, etc. into a common analysis framework

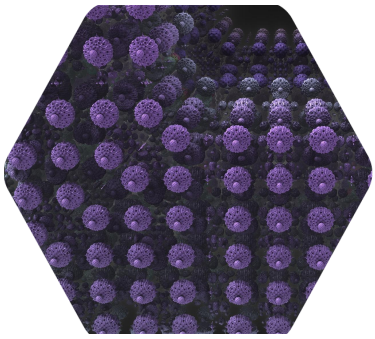
Image credits:
https://pixabay.com/en/kingfisher-bird-alcedo-atthis-1068480/ (cc0)
https://pixabay.com/en/rotterdam-cycling-bicycle-rental-1199442/ (cc0)
https://pixabay.com/en/egg-food-many-duplicate-easter-85641/ (cc0)
https://pixabay.com/en/female-diary-write-beautiful-865110/ (cc0)
https://pixabay.com/en/abstract-blurred-blur-color-mix-859315/ (cc0)

Ways to build effective ML

Big Data



Feature Engineering



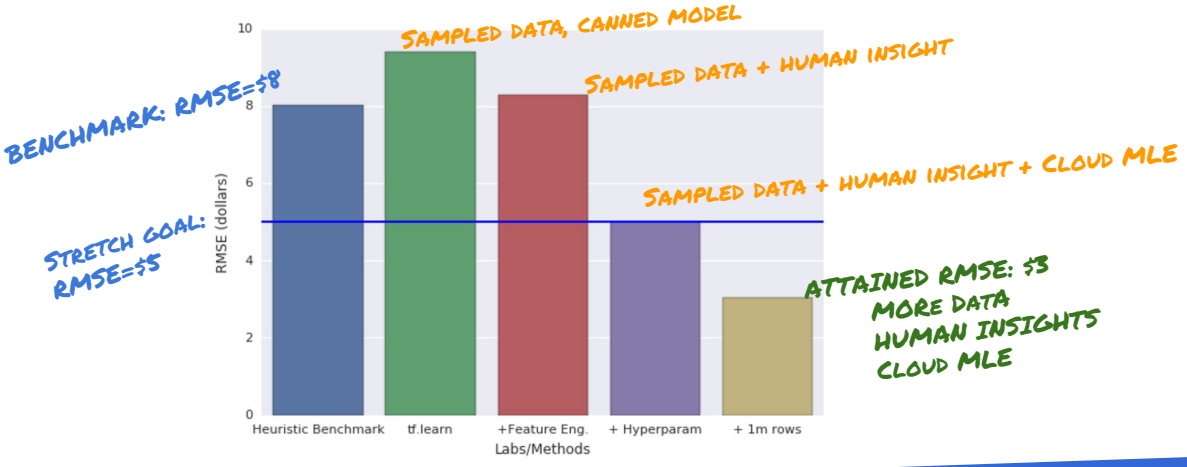
Model Architectures

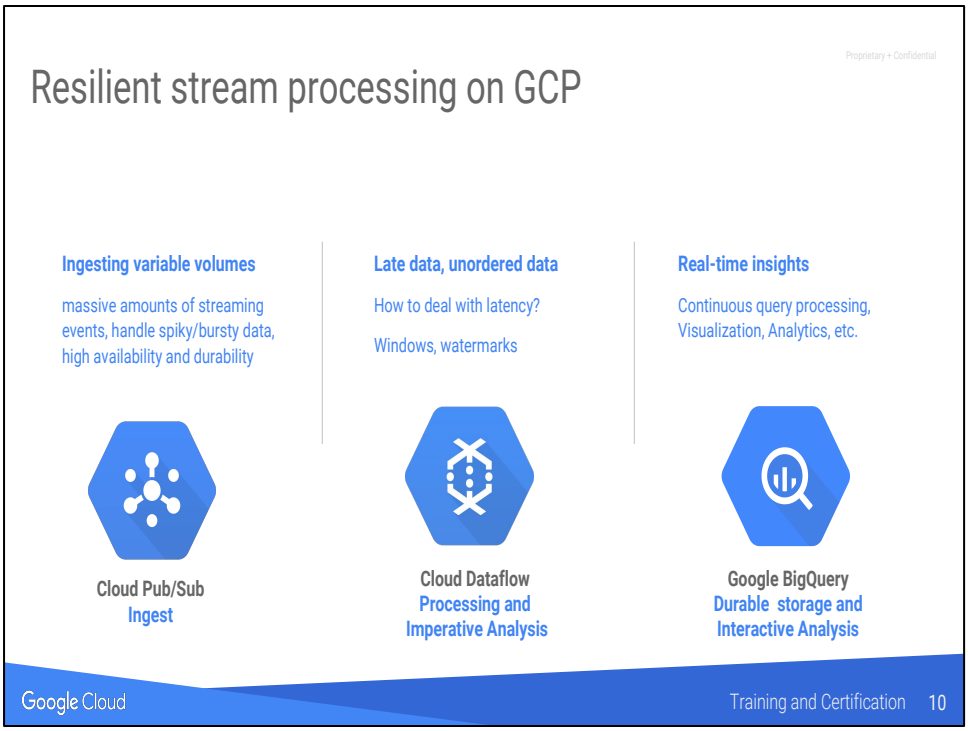
Notes:

https://pixabay.com/en/large-data-dataset-word-895563/ (cc0)
https://pixabay.com/en/fractal-complexity-render-3d-1232494/ (cc0)
https://pixabay.com/en/robot-artificial-intelligence-woman-507811/ (cc0)

Now that you know *how* to build ML, let's learn how to do it well in the rest of the course.
Ordered from easiest to most difficult.

Accuracy improves through feature engineering, hyperparameter tuning, and lots of data





Notes:

PubSub is your global message bus.

Dataflow is capable of doing batch and streaming ….code does not change.

BigQuery gives you power of analytics.

BigTable when you get overwhelming volume.

Next steps: Google-certified Data Engineer

GOOGLE CERTIFIED PROFESSIONAL

Data Engineer

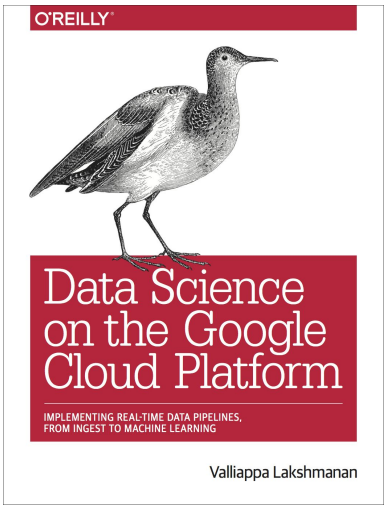
A Google Certified Professional - Data Engineer enables data-driven decision making by collecting, transforming, and visualizing data. The Data Engineer designs, builds, maintains, and troubleshoots data processing systems with a particular emphasis on the security, reliability, fault-tolerance, scalability, fidelity, and efficiency of such systems.

The Data Engineer also analyzes data to gain insight into business outcomes, builds statistical models to support decision-making, and creates machine learning models to automate and simplify key business processes.

https://cloud.google.com/certification/data-engineer

Resources

- Big data and machine learning blog
  https://cloud.google.com/blog/big-data/
- Google Cloud Platform blog
  https://cloudplatform.googleblog.com/
- Data Science on the Google Cloud Platform
  http://shop.oreilly.com/product/0636920057628.do



Notes:

Permission to use book cover image in slides and blog posts was granted by O'Reilly. Lak forwarded the email from O'reilly to cloud-legal@ on March 1, 2017.

Please take 5 minutes to give us feedback

g.co/CloudTrainEval

This URL is case-insensitive

The class code is always a tag in the QL class. Instructor can always find the class code (if they don't have it already) when they are in the class in QL, click Edit Class and the code is in the tags. Not ideal - but we'll figure out how to make it more prominent


cloud.google.com
Image by Connie Zhou