Big Data

Platform and Architecture Recommendation

Sofyan Hadi Ahmad
sofyan.ahmad@mandalalabs.com

November 2018

Why Big Data is Important?

Despite the hype, many organizations don’t realize they have a big data problem or they simply don’t think of it in terms of big data. In general, an organization is likely to benefit from big data technologies when existing databases and applications can no longer scale to support sudden increases in volume, variety, and velocity of data.

Failure to correctly address big data challenges can result in escalating costs, as well as reduced productivity and competitiveness. On the other hand, a sound big data strategy can help organizations reduce costs and gain operational efficiencies by migrating heavy existing workloads to big data technologies; as well as deploying new applications to capitalize on new opportunities.

How Does Big Data Work?

With new tools that address the entire data management cycle, big data technologies make it technically and economically feasible, not only to collect and store larger datasets, but also to analyze them in order to uncover new and valuable insights. In most cases, big data processing involves a common data flow – from collection of raw data to consumption of actionable information.

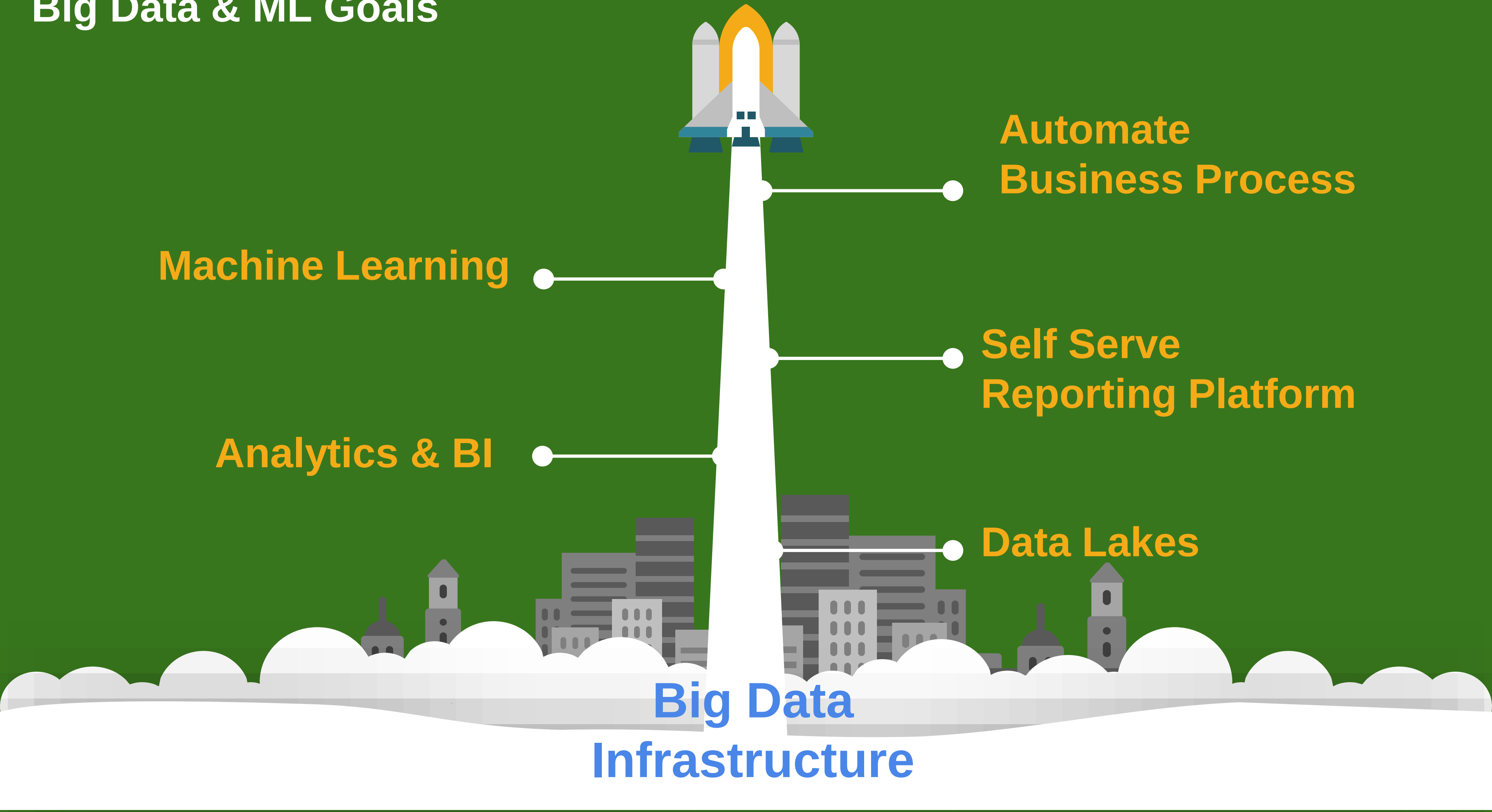
- **Collect**. Collecting the raw data – transactions, logs, mobile devices and more – is the first challenge many organizations face when dealing with big data. A good big data platform makes this step easier, allowing developers to ingest a wide variety of data – from structured to unstructured – at any speed – from real-time to batch.

- **Store**. Any big data platform needs a secure, scalable, and durable repository to store data prior or even after processing tasks. Depending on your specific requirements, you may also need temporary stores for data in-transit.

- **Process & Analyze**. This is the step where data is transformed from its raw state into a consumable format – usually by means of sorting, aggregating, joining and even performing more advanced functions and algorithms. The resulting data sets are then stored for further processing or made available for consumption via business intelligence and data visualization tools.

- **Consume & Visualize**. Big data is all about getting high value, actionable insights from your data assets. Ideally, data is made available to stakeholders through self-service business intelligence and agile data visualization tools that allow for fast and easy exploration of datasets. Depending on the type of analytics, end-users may also consume the resulting data in the form of statistical “predictions” – in the case of predictive analytics – or recommended actions – in the case of prescriptive analytics.

Big Data & ML Goals



Automate Business Process

Machine Learning

Self Serve Reporting Platform

Analytics & BI

Data Lakes

Big Data Infrastructure

Platform

Which platform is the best to implement big data?

- Cloud or on-Premise?
- If Cloud, Google, AWS, Azure, or Cloudera Cloud?
- If on-premise, is Cloudera is the best way to go?

Cloud or On-Premise?

- Why Cloud
  - Planning, research, analytics for a lower
  - No Server
  - Get Started Tomorrow
  - Grow as fast you can imagine
  - Get out of old IT business line
  - Part of your Opex, not Capex
  - Low commitment, you can jump from one to another provider, just like counting 1, 2, 3
  - Safer and 99% uptime
- Why On-Premise
  - Meet your inhouse policies
  - You own the software, and the hardware

Recommendation: Cloud

Why pick cloud?

1. Flexibility. Cloud-based services are ideal for businesses with growing or fluctuating bandwidth demands.
2. Disaster recovery. Businesses of all sizes should be investing in robust disaster recovery, by using cloud, you avoid large up-front investment and roll up third-party expertise as part of the deal.
3. Automatic software updates. Provider take care that for you and roll-out regular software updates – including security updates – so you don’t have to worry about wasting time maintaining the system yourself.
4. Capital-expenditure Free. You simply pay as you go and enjoy a subscription-based model that’s kind to your cash.
5. Increased collaboration. When your teams can access, edit and share documents anytime.
6. Work from anywhere. With cloud computing, if you’ve got an internet connection you can be at work. And with most serious cloud services offering mobile apps, you’re not restricted by which device you’ve got to hand.
   The result? Businesses can offer more flexible working perks to employees so they can enjoy the work-life balance that suits them – without productivity taking a hit.

1. Document control. Before the cloud, workers had to send files back and forth as email attachments to be worked on by one user at a time. Sooner or later – usually sooner – you end up with a mess of conflicting file content, formats and titles.
2. Security. Lost laptops are a billion dollar business problem. And potentially greater than the loss of an expensive piece of kit is the loss of the sensitive data inside it. Cloud computing gives you greater security when this happens.
3. Competitiveness. Wish there was a simple step you could take to become more competitive? Moving to the cloud gives access to enterprise-class technology, for everyone.
4. Environmentally friendly. While the above points spell out the benefits of cloud computing for your business, moving to the cloud isn’t an entirely selfish act. The environment gets a little love too. When your cloud needs fluctuate, your server capacity scales up and down to fit.

Any three of the above benefits would be enough to convince many businesses to move their business into the cloud.
But when you add up all ten? It’s approaching no-brainer territory.

Google, AWS, Azure, or Cloudera Cloud?

Why Google

- **Big Data, no waiting.**
  Based on the same distributed data services we use at Google, Google BigQuery, Google Cloud Datalab and Google Cloud Dataproc are changing how you analyze and use data. Customers say tools like BigQuery are “nearly magical” because of their performance. Queries that used to take hours or days now take minutes or seconds. The result: more insights and value, realized by more people in more companies. Learn more
- **Context-rich applications**
  Good apps tailor their behavior to us. Great apps delight us by suggesting what we want before we know it ourselves. Apps should respond to context, telling us the best route to drive right now, not just when we started. Products like Google Cloud Dataflow and Google Cloud Pub/Sub make it easier for your code to use huge amounts of data to deliver amazing context-rich experiences.
- **Citizen Data Science**
  Rather than keep your most powerful data tools in the hands of a few experts, Google Cloud Platform unlocks data to empower your entire organization. Our tools like BigQuery and Cloud Datalab bring data directly to the people who run your business, because they’re most likely to find the insights that create value.
- **Machine Intelligence**
  Google Cloud Machine Learning gives everyone access to the deep learning systems that power services like Google Translate, Google Photos, voice search in the Google app, and smart replies in Gmail. Soon available as a cloud service (alpha), so it’s easy to incorporate in your app.

Why AWS

- **AWS is the cloud market leader**
  AWS is a market leader with 40% of market capitalization in 2017. Its vast list of features makes it lucrative for larger corporations whereas a growing infrastructure promises scalability and price cuts for upstarts.
- **Very Flexible**
  AWS enables you to select the operating system, programming language, web application platform, database, and other services you need. With AWS, you receive a virtual environment that lets you load the software and services your application requires. This eases the migration process for existing applications while preserving options for building new solutions.

Why Azure (ADLA: Azure Data Lake Analytics)

- **Only one language to learn**
  OK, this is kind of tricky, because, although there is only one language to learn, U-SQL, it's really an amalgam of SQL and C# – so if you're familiar with either or both of those languages then you'll be ready to tackle U-SQL. If you're not familiar with either, then it's OK, you don't have to be a SQL or C# master to understand U-SQL. With Hadoop, you'll have a few more languages to learn – I'd say at least six (Hive, Pig, Java, Scala, Python, Bash; yup, six).
- **Only offered as a platform service**
  Hadoop comes in many different flavors, some running on-premises, others running in the cloud. Some are managed BY you, others are managed FOR you. ADLA, however, is offered ONLY as a platform service in the Microsoft Azure Cloud. It's managed by Microsoft — you'll never have to troubleshoot a cluster problem with ADLA (only your own code). It's also integrated with Azure Active Directory, so you don't have to manage security separately.
  Really, all you have to worry about when you set up an Azure Data Lake Analytics account is building your application.
- **Pricing per job; not per hour**
  Most Big Data cloud offerings that are available are priced per hour based on how long you keep your cluster up and running. ADLA takes a different approach to pricing. With ADLA, you pay for each individual job that is run. As a matter of fact, just owning an Azure Data Lake Analytics account doesn't cost anything. You aren't even billed for the account until you run a job. That's pretty unique in the Big Data space.

Why Cloudera Cloud

- **Backed by best Big Data Software**
  No one knows Apache Hadoop like Cloudera. Period!.
- **Elastic**
  Cloudera’s platform can leverage elastic infrastructure to size compute and storage independently, grow and shrink clusters dynamically, and clone net-new clusters for ad-hoc, transient workloads.
- **Portable**
  Preserve business flexibility and minimize cloud lock-in. Run your Cloudera workloads on any public cloud, or even in multi-cloud environments. You can even switch from Cloud to On Premise without hassle
- **Enterprise Grade**
  Reduce your risk with the only big data platform that delivers comprehensive manageability, availability, security, and governance required for production workloads.

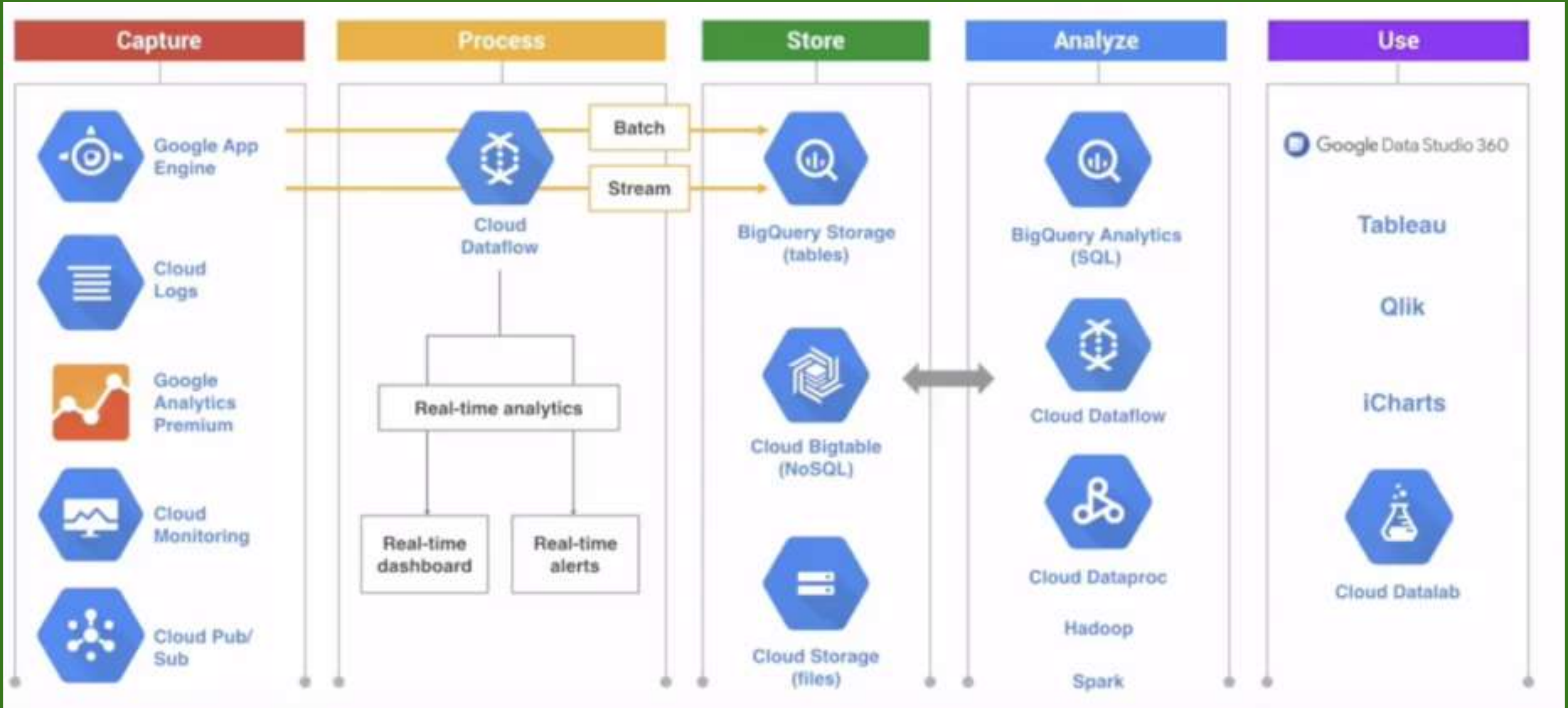
Recommendation: Google Cloud Platform

Benchmarks:

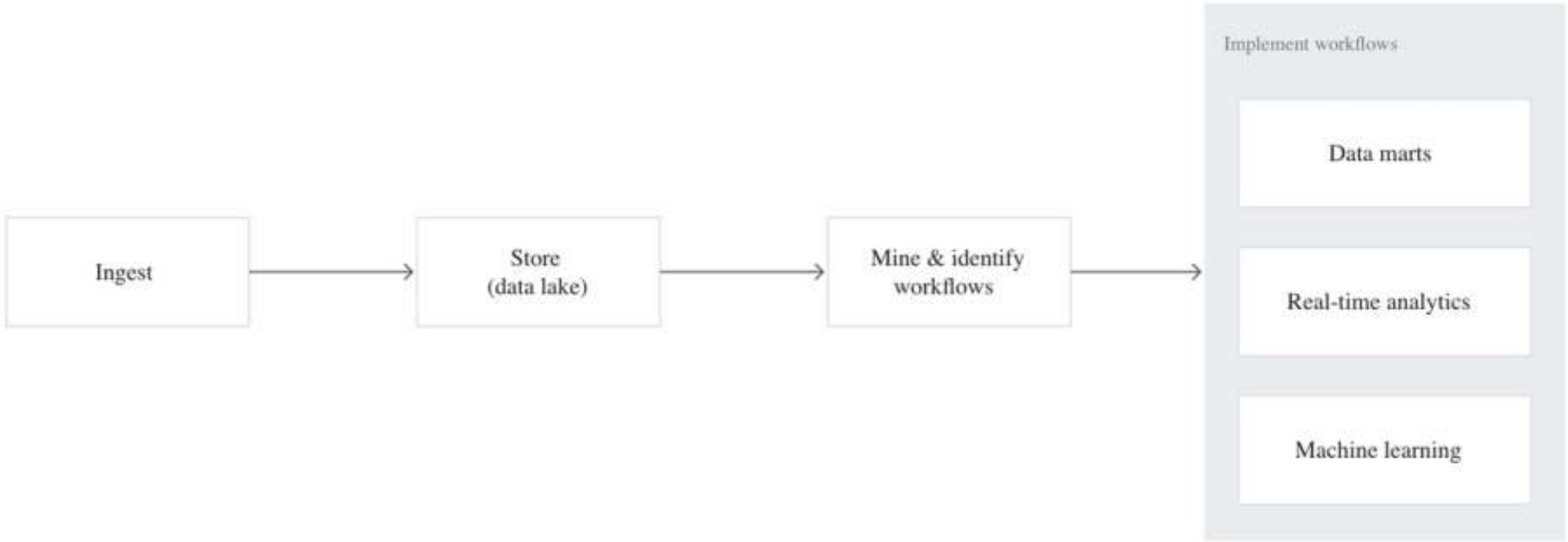
- Tools and Support: Google
- Platform Expertise: Google
- Performance: Google
- Integration: Google
- Flexibility: Cloudera
- Easy To Learn: AWS
- Pricing: Azure

Architecture

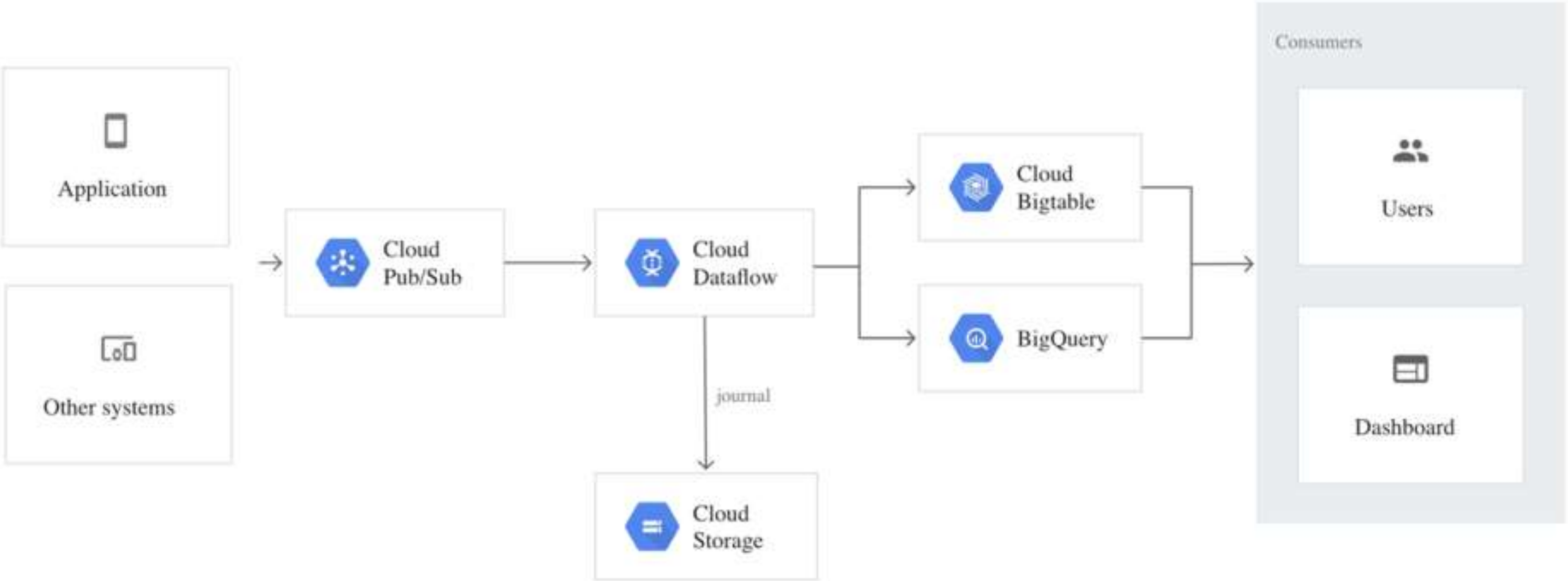
GCP: Big Data Life Cycle



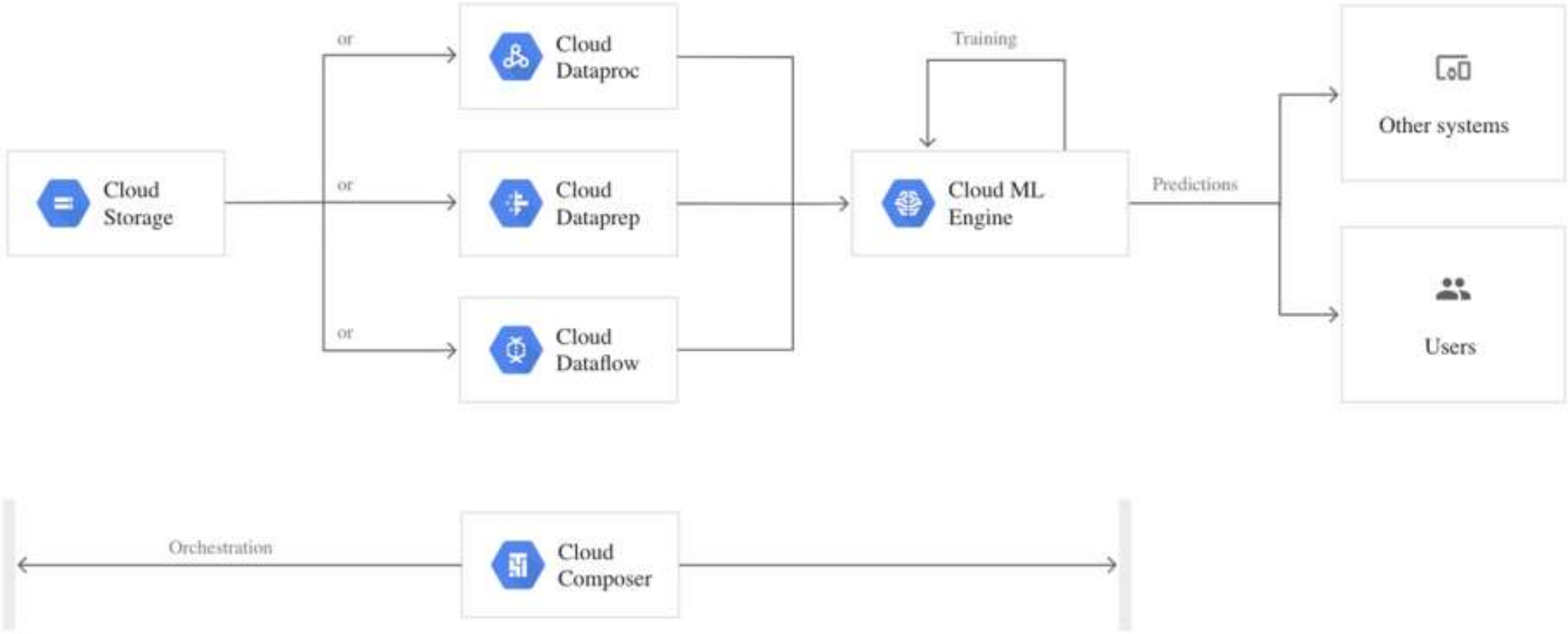
Basic Data Lake Workflows



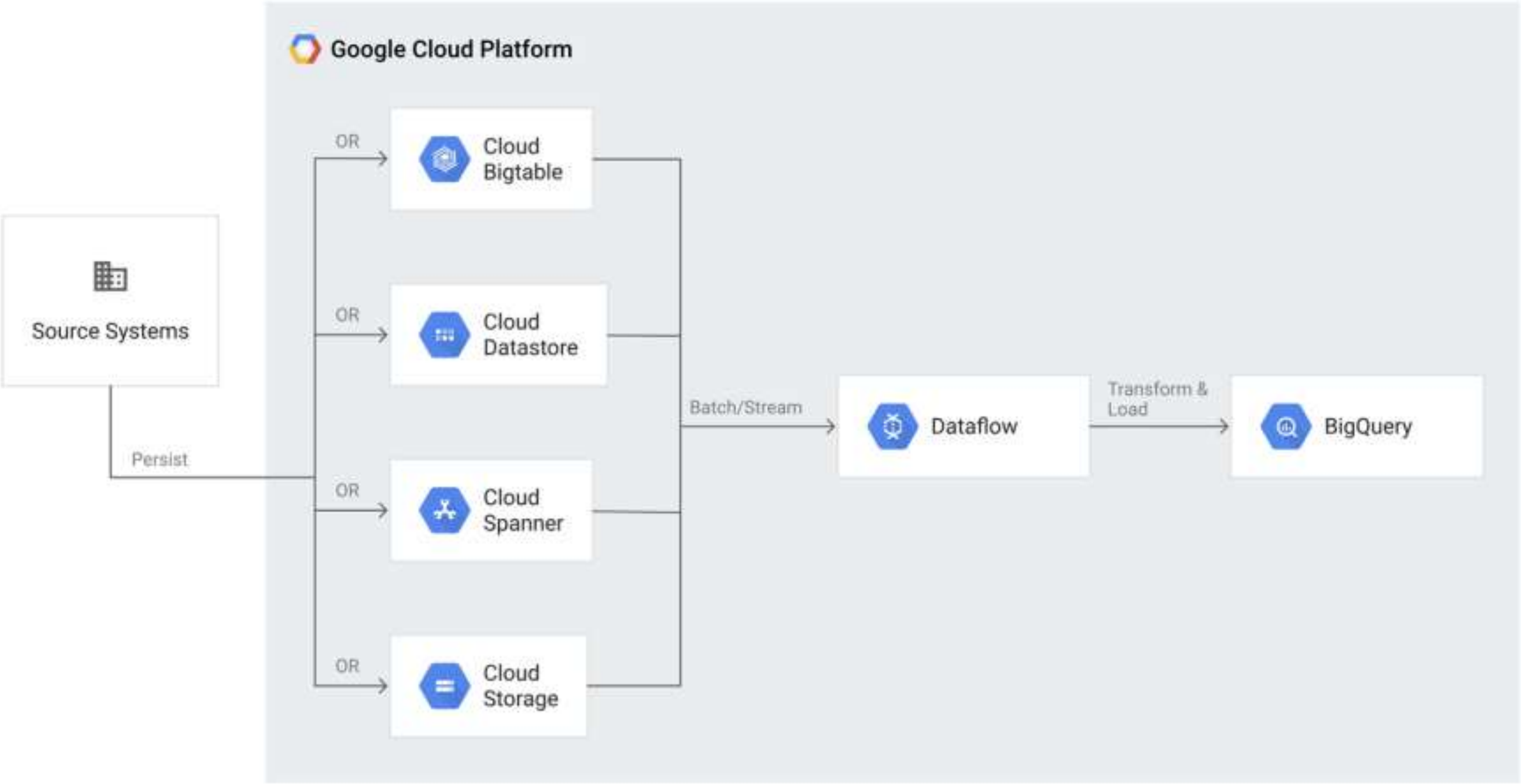
Real Time Self Reporting Big Data Dashboard



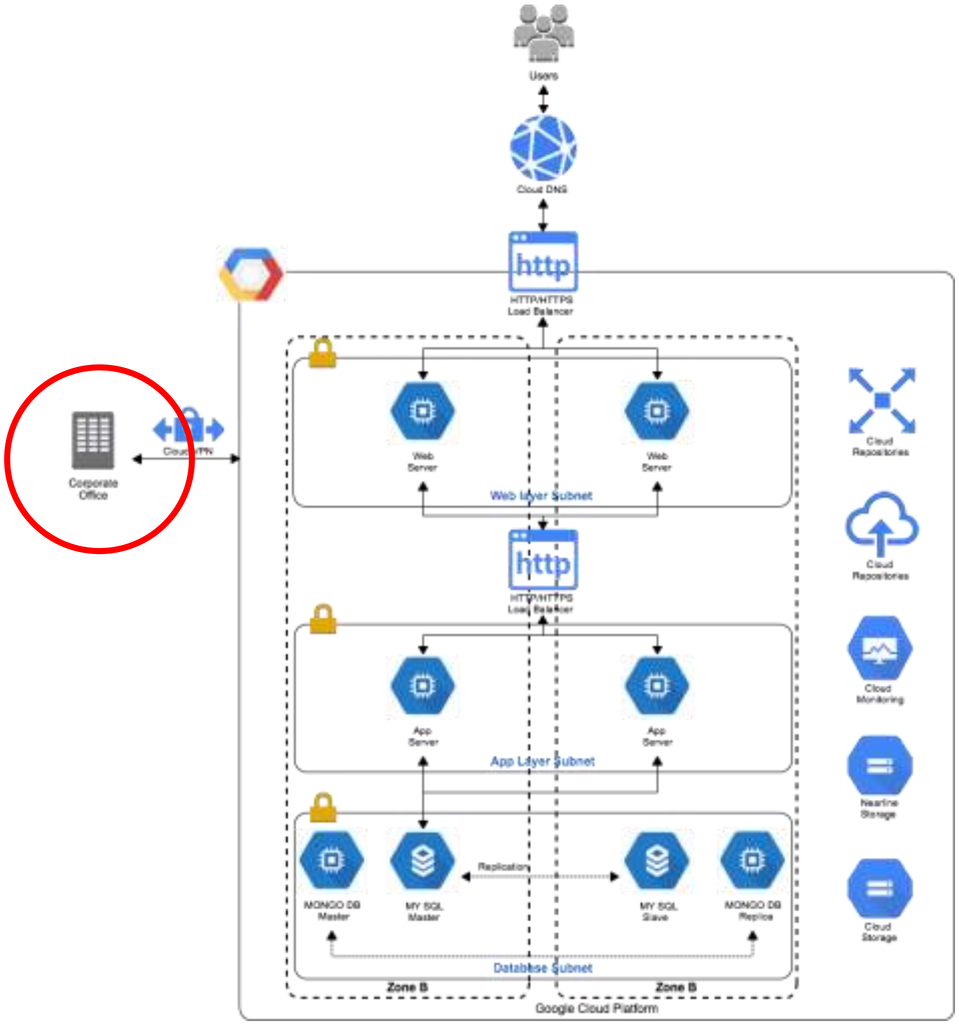
Machine Learning Integration



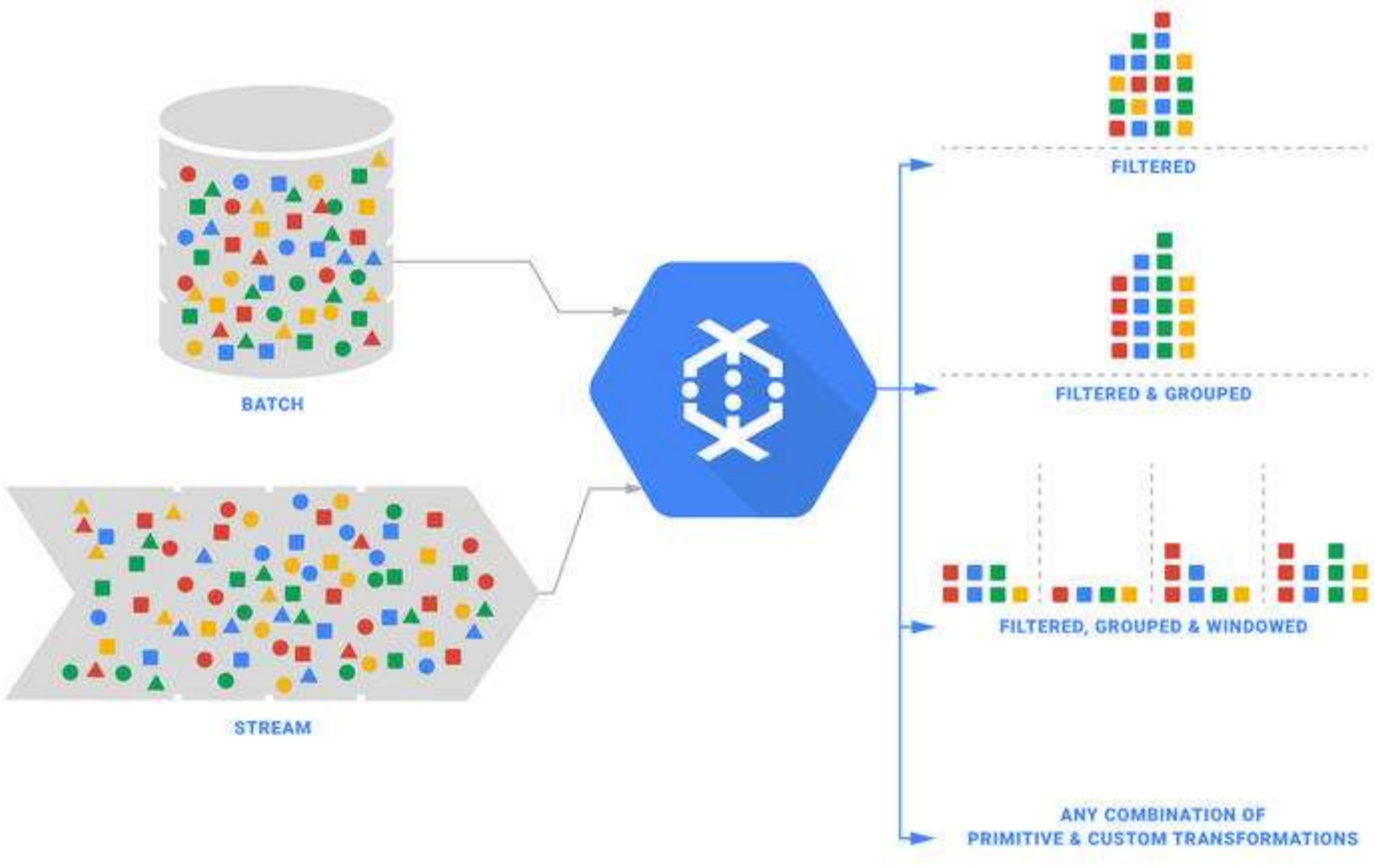
Data Warehouse transitions to GCP



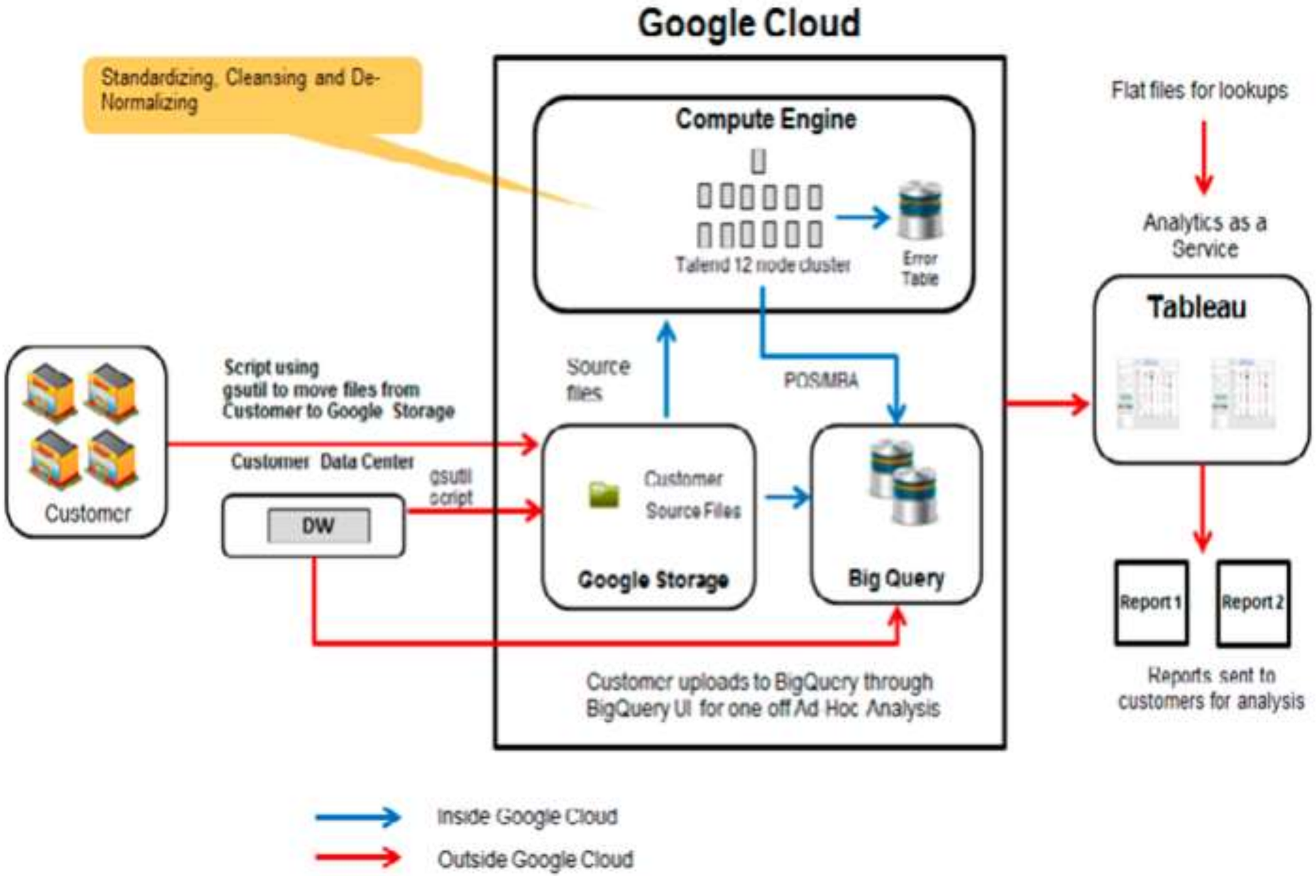
How to run GCP and Existing System On-Premise?



Understanding data stream using Data Flow Engine


BATCH
STREAM
FILTERED
FILTERED & GROUPED
FILTERED, GROUPED & WINDOWED
ANY COMBINATION OF
PRIMITIVE & CUSTOM TRANSFORMATIONS

Ingest Data into Google BigQuery Using Talend



Standardizing, Cleansing and De-Normalizing

Script using gsutil to move files from Customer to Google Storage

Customer uploads to BigQuery through BigQuery UI for one off Ad Hoc Analysis

Reports sent to customers for analysis

Inside Google Cloud

Outside Google Cloud

Question?

sofyan.h.ahmad@gmail.com