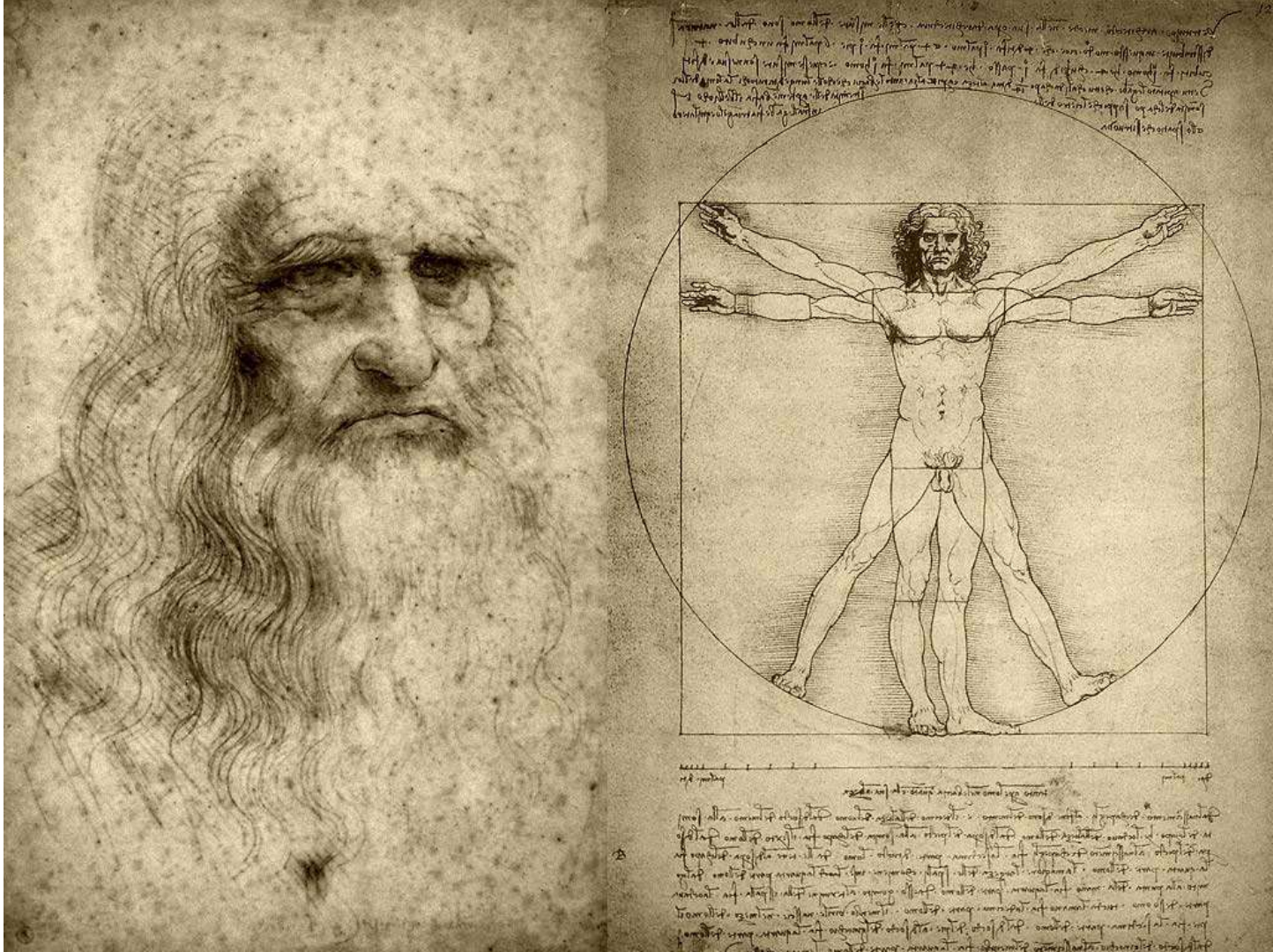
GCP Big Data Demystified #1 | Investing.com Big Data Journey

Omid Vahdaty,
Big Data Ninja

Welcome

Big Data Demystified Meetup

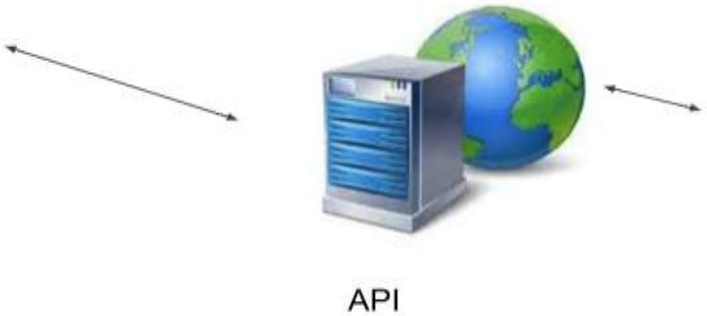


Disclaimer

- I am not the best, I simply love what I do VERY much.
- You are more than welcome to challenge me or anything I have to say as I could be wrong.

A long time ago
in a galaxy far, far away....

In the Past(web,api, ops db, data warehouse)



Then came Big Data...

Then came the cloud...

Then came the invoice ...

It keeps growing....

Solution?

Cloud

+

Big Data

=

Data Engineering

Today’s use case? Investing.com

Part1

What is the use case?

Investing.com use case

- Top X financial Publisher, **Fastest growing** financial portal
- +300 Employee
- New **Data Labs departmen**t
- 1.2 Billion monthly events
- 3 **Datacenter** Globally distributed ±1000 servers
- 7+ **data teams** with different **data need**s
  - Management
  - Madrid
  - BI team
  - DE team
  - DS team
  - Analysis team
  - Product Data
  - Global offices around the world. Including daughter companies

More details on the use case

Data sources :

- Internal Ad Server (AKA Krank) - onsite.
- GA360
- DFP
- Operational DB: MySQL - onsite.
- Data Warehouse: MySQL - onsite.
- ± hundreds of API data sources to sync with

Data Volume: +20 TB and growing rapidly

Data Velocity: 1.2B events monthly

Data Variety: Json, CSV, XML, API

Data Veracity: major bugs in the data found monthly .

Part2

Where should we build a data platform?

Step 2: Where to build the data platform?

amazon web services

Google Cloud Platform

Data Center VS GCP VS AWS | Pros

Data Center Pros

- We already have servers
- Data is already there
- No need to rewrite hunderded of ETLS
- No need to educate an entire org to cloud thinking

AWS Pros

- Wide range of Big data features(Athena/Spectrum/Redshift/Glue)
- No Learning Curve (not my first Rodeo)
- Strong Hadoop distribution: EMR
- Account Segregation

amazon web services

GCP Pros

- Big Query | PAYG
- GA data already inside
- 400$ credit from GA
- No Ingest Costs
- No Transformation cost
- No Export Costs

Google Cloud Platform

Data Center VS GCP VS AWS | Cons

Data Center Cons

- Dependency on OPS team
- Change is slow

AWS Cons

- **Migrating data effort** from GCP
- **Cost** of migrating data
- wasted GCP Credits
- Training an entire org.

amazon web services

GCP Cons

- My Learning Curve
- External Table
- Very easy to make

  costly mistakes in BQ

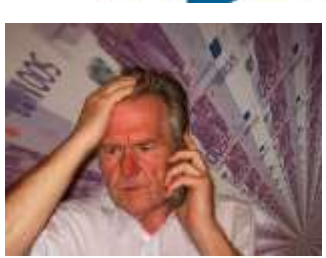
- Training an entire org.
- **UTF16** BQ not supported

Google Cloud Platform

Part3

mapping

Now days…



Step 3: Mapping

- Get the list of all the data sources
  - **Volume**
  - **Velocity**
  - **Access patterns** from different **Business Units**
- Get the list of all the data technologies **relevant** to your **architecture**
  - Understand the **performance** limitations
  - Understand the **limits** of the products
  - Understand **cost** model
  - Understand the **security**

Step 3.1 : which technologies need to be deprecated/discarded in the organization and why

- **MySQL** as a DWH
  - No analytical functions
  - Row based
  - Not easily scalable on DC
- **Windows** , not a good match for DE and DS ecosystem.
- **Orchestration** - should we keep using talend? Or switch to something more modern like managed airflow (cloud agnostic, “free”)?

Part4

Architecture

Faster, Cheaper, Simpler

"Everything should be made as simple as possible. But not simpler."

-Albert Einstein

Step 4: Gather requirements & Design the Architecture

- **Requirements**
  - Keep an **Open mind**....
  - **Time to market**.
  - Match the technology to the use case and not the other way around.
  - **Segregations** of teams and their budget and **data governance** (who has has access to what)
  - Get monthly **budget approved**.
  - Generate Quick win without breaking existing data workflow
- **Architecture**
    - **Decoupled → so you can adjust yourself to changing times.**
    - **PaaS → focus on the business not on the technology.**
    - **Faster , Cheaper, Simpler → this is data data engineering in a nutshell.**

Step 4.1 Which technology was not selected

- **Dataproc**
  - Missing PaaS features (compared to EMR)
  - No GCP Support on missing above PaaS features
  - We didn't know exactly what are our monthly **data scan needs**. It sounded like an overkill for the first phase.
- **DataFlow**
  - Already had talent :(
  - We dont have streaming, unlikely to have in the next 2 years.
  - Requires Steep learning curve + coding skills (unless using templates wizards)
- **BigQuery And external table**
  - No export to parquet
  - No flexibility in the create table to change null marker
  - Cost plugin, can’t estimate cost on external table
  - **BQ FLAT rate** was discarded until we have some metrics on how data scan monthly.

4.2 which technologies did we end up with and why?

- **Big Query - AD HOC querying, Interactive mode**
  - Get query metrics. Unknown volume? queries/sec? **monthly data scan**?
  - **Slowly** adjust the **organization to adapt** it way of thinking to: Faster, Cheaper,Simpler.
  - **Managed service**
  - **Free import/export**
- **Data Labs**
  - **Time to market**
  - Simple
  - Managed service
- **GCS**
- **Pub/Sub**
  - In the near future for messaging.
  - Unknown **ingress rate**
  - Managed service

Step 4.2 Selected technologies and their issues

- BigQuery
  - SQL missing:
    - **Load CSV** requires **cli command**
    - **Show create table**
    - **Create table is annoying.**
  - **cost plugin**
  - **Cross region** data set joins are not supported.
  - **No delete policy** - custom role
  - **One truth** - sharing dataset across multiple projects
  - Missing **Quota per per Group / person** :(
- Data labs
  - **Single user** environment
  - **Admin user VS DS user**
- GCS - **Cross project** bucket sharing.

PayAsYouGo RATE VS FLAT

- Good when there **unknown variables in you future plan**
- Maximum **performance**
- Changes **ORG state of mind** to keep **cost** in mind.

- **Predictable** cost
- **Cheaper** than PAYG

Part5

Communicate

Faster, Cheaper, Simpler

"Everything should be made as simple as possible. But not simpler."

-Albert Einstein

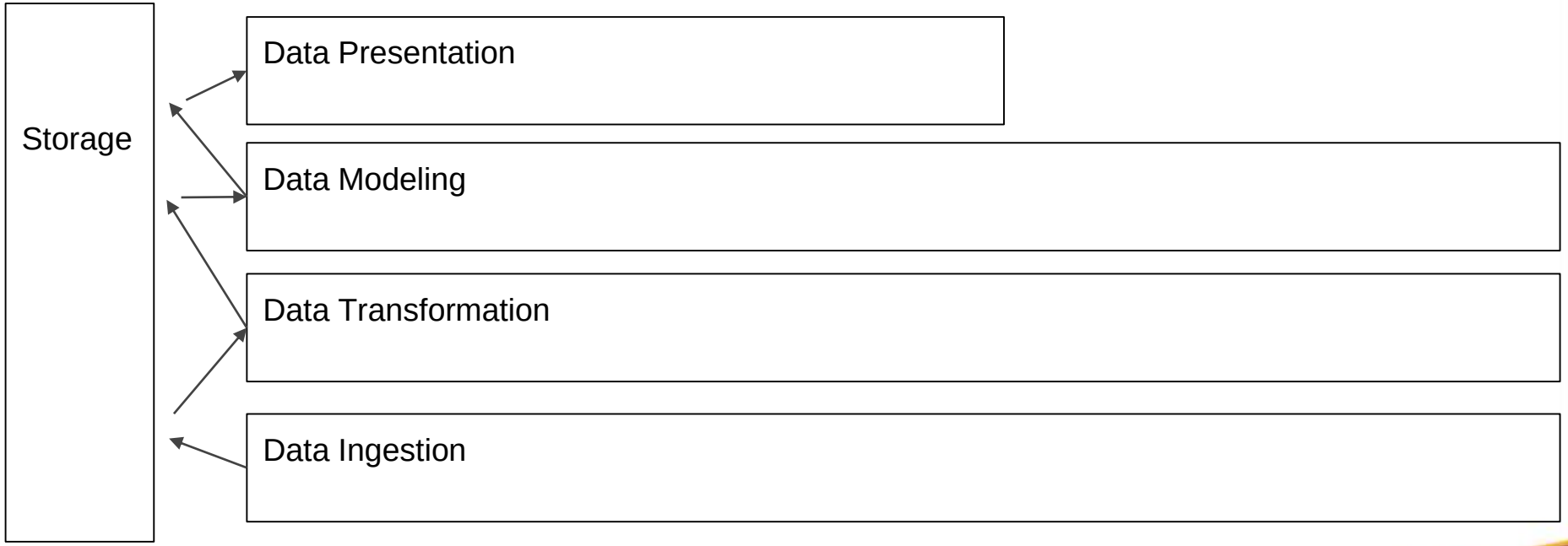
Step 5: communicate Your Plan

- Management team
- Business units
- Team members
- Basically every stake holder you can think of

Jargon

Columnar VS Row Based
External table VS local Table
Storage VS Compute
Data Ingestion
Data Transformation
Data Modeling
Data Presentation

ARCHITECTURE | General



ARCHITECTURE | general

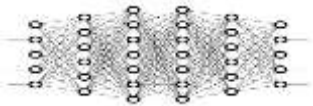
Google Cloud Storage

Parquet

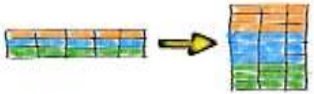
CSV,
RAW

Data **Visualization**

Data **Modeling** (joins/agg/ML/R)

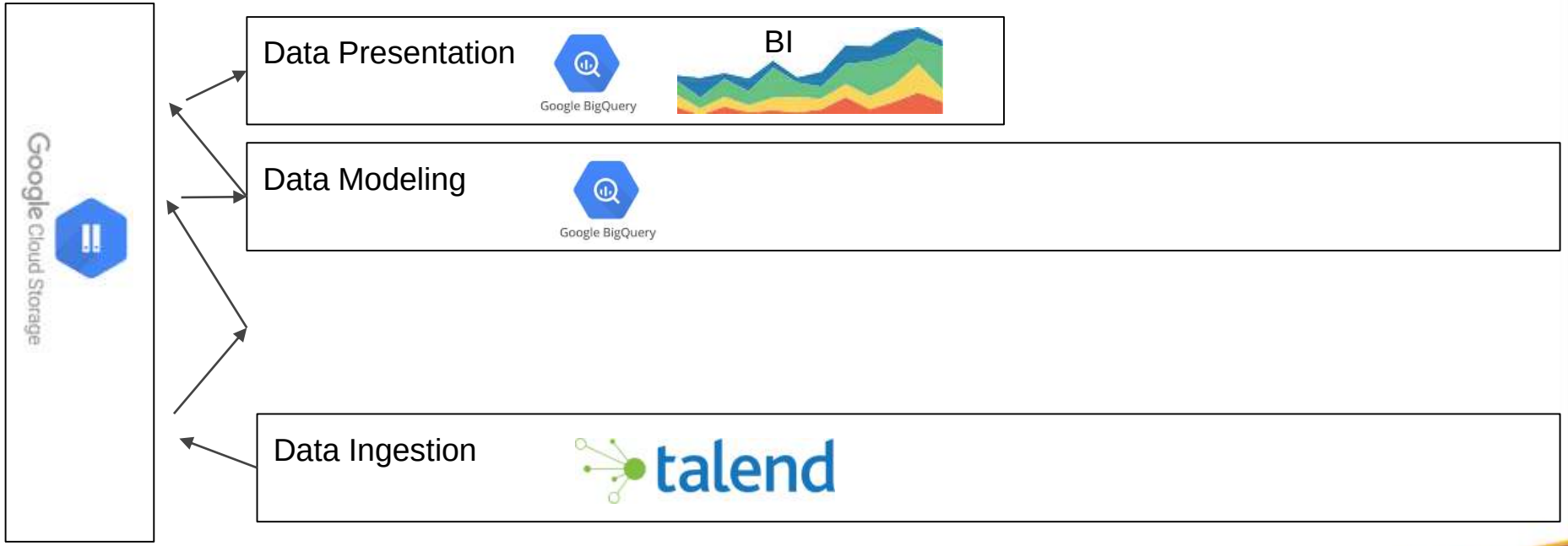


Data **Transformation** (**row** to **colunar** + cleansing)

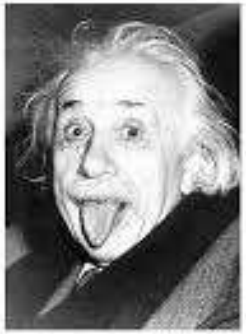


Data Ingestion
(file based ETL from remote DC)

ARCHITECTURE | Quick Win



Architecture | Faster? Cheaper? Simple?



"Everything should be made as simple as possible. But not simpler."

-Albert Einstein

Ingestion Guidelines

Keep the data in the following format:

1. **Parent folder** for each table inside
   a. each parent folder in a **sub folder** called “dt=today” . e.g.: “dt=2018-11-01”
   b. in side each DT folder there will be file per day = partitioning, the name will be table_name_2018-11-01.gz
   c. save it on google cloud storage, gzipped.
   d. replace a whole file when needed. :)

Transformation

- Is all about switching from CSV to columnar data structures
- In first stage - will be done by BigQuery
- Now - Big Query - internal table.
- Future - Hadoop - External tables on GCS

Modeling

- via Bigquery
- Modeling the data is a matter of requirements from all Business units. the idea is to **model the data in matter that the queries will be simple and readable**.

Big Query

- Very fast
- Very Powerful
- Highly scalable
- Columnar

- Very Expensive
- Very easy to make 10000$ mistakes.
- Requires learning curve.
- requires to load data

Recommendation: Quota

- currently each user can process upto 39TB per day.
- The entire team can process up to 40TB per day
- Quota can according to Business unit budget by divided to different projects.

Recommendation: BQ MATE

New Query

```
SELECT * FROM `123492829.ga_sessions_*`
WHERE _TABLE_SUFFIX BETWEEN '20181001' AND '20181020'
```

Valid: This query will process 662 GB when run. The cost will be around $3.31

Standard SQL Dialect

RUN QUERY | Save Query | Save View | Format Query | Schedule Query | Show Options | BQ Mate | Query failed

Results | Details

Query Failed

Error: Custom quota exceeded: Your usage exceeded the custom quota for QueryUsagePerUserPerDay, which is set by your administrator. For more information, see https://cloud.google.com/bigquery/cost-controls

Job ID: gap---all-sites-1245:US.bquijob_5299d4b5_166e9d54257

```
SELECT * FROM `123492829.ga_sessions_*`
WHERE _TABLE_SUFFIX BETWEEN '20181001' AND '20181020'
```

BigQuery Mate

Super Query @ Investing

Google Cloud Platform
GAP - All Sites
BigQuery
superQuery

superQuery
Board 2
VISUALIZE
EDITOR

PROJECTS
SCHEMA
BOARDS

Search tables, datasets, projects

PROJECTS:
- datascience-02122018
- investing-analysis
- investing-media
- madrid-investing

GITHUB

fromDate: Today
Add Variable

Compose Query | Tab 1 | Tab 2 | Tab 3 | Tab 4 | Tab 5

1

Options
SQL Dialect: Auto-detect
Select Project: datascience-02122018
Caching: SuperQuery
Query Priority: Interactive

superQuery
$0.00
0 Byte

Project ID: Datascience-02122018

Results
Details

No data to show

Part6

Challenges?

Faster, Cheaper, Simpler

"Everything should be made as simple as possible. But not simpler."

-Albert Einstein

Challenges

- **Build as you go** - live system, legacy ETLS
- **Data silos** vs. One truth
- **Re-org** of data teams as we go.
- Workarounds with bigQuery (only with **GA 360 session table** - poorly modeled by GA team)
- Decoupling dashboards from data (joins on presentation layer)
- Production data ETL latency VS streaming
- Communicating change throughout large global organizations.

Part7

Dream

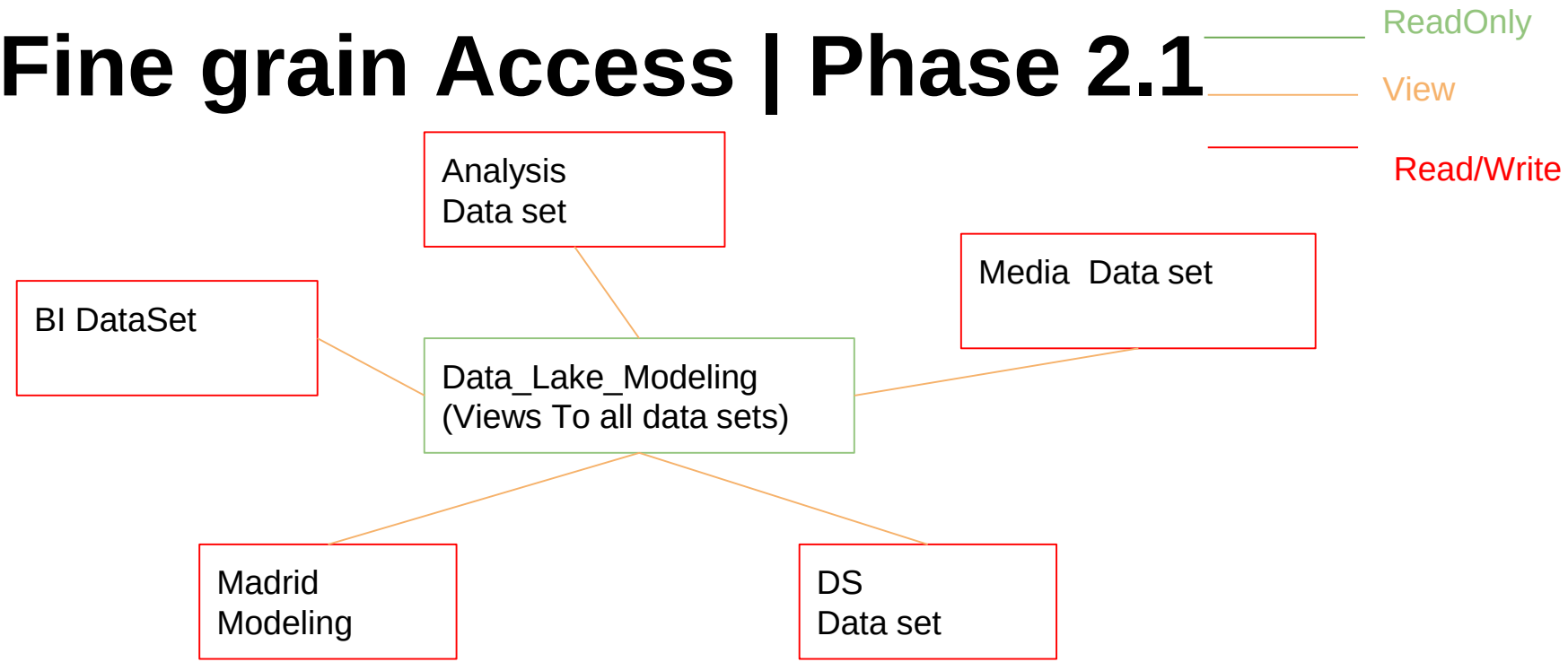
Faster, Cheaper, Simpler

"Everything should be made as simple as possible. But not simpler."

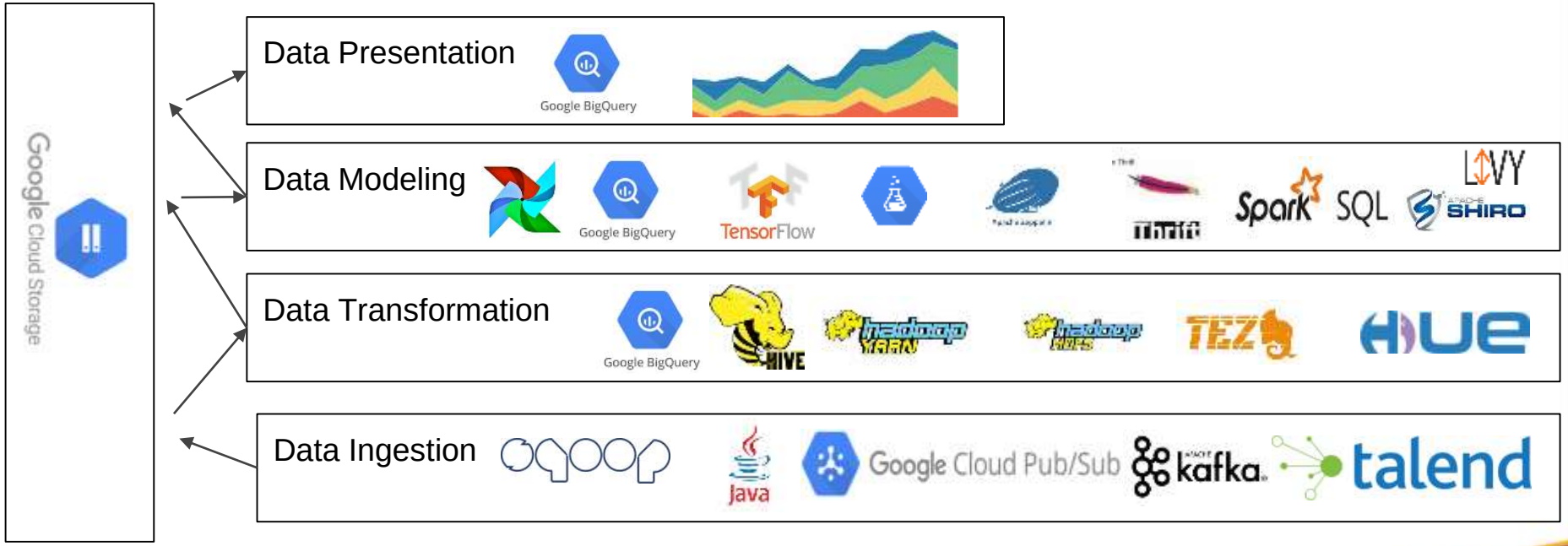
-Albert Einstein

Quick word
about the future...

Fine grain Access | Phase 2.1



ARCHITECTURE | Future



Summery

And
Call to Actions

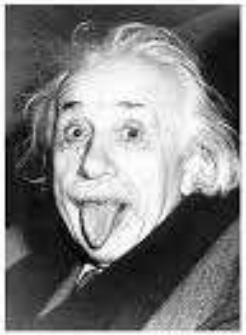
Steps to build big data architecture

1. Use Case?
2. Where ?
3. Mapping?
4. Architecture?
5. Communicate?
6. Challenges?
7. Dream!

Summary… Data Engineering is all about:

Faster

Cheaper



"Everything should be made as simple as possible. But not simpler."

-Albert Einstein

Simpler

How to get started | Call for Action

Lectures: AWS Big data demystified lectures #1 until #4

[AWS Big Data Demystified Meetup]

[Big Data Demystified meetup]

Stay in touch...

- Omid Vahdaty
- +972-54-2384178
- https://big-data-demystified.ninja/
- Join our meetup, subscribe to youtube channels
  - https://www.meetup.com/AWS-Big-Data-Demystified/
  - https://www.meetup.com/Big-Data-Demystified/
  - Big Data Demystified YouTube
  - AWS Big Data Demystified YouTube
  - WhatsApp group