

# Day in the life of “Data Engineer”

| July 11, 2018  
Rashmi Shamprasad



# Agenda.

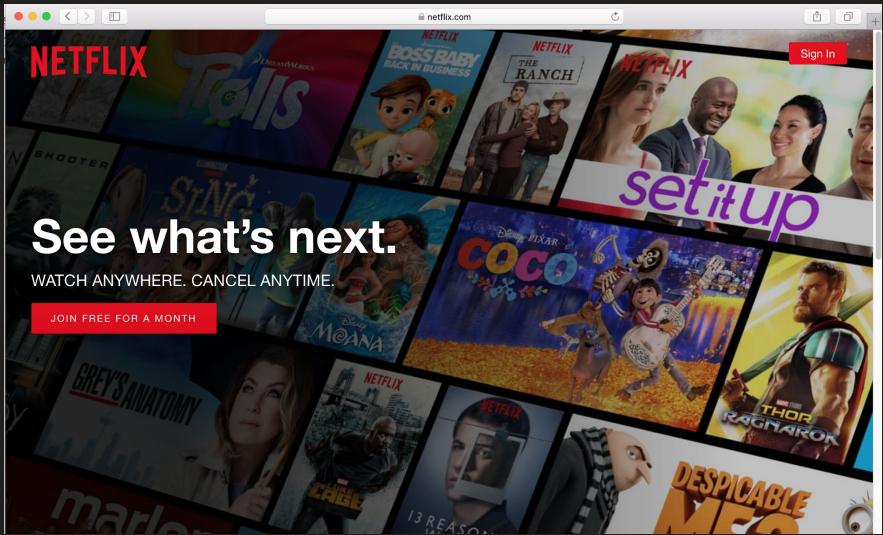
- A Typical Data Engineering Project
- Hands on Exercises

# A Typical Data Engineering Project



# Starts with a Problem Statement ...

e.g. How many visitors does Netflix get every day ?



# Problem Statement - How many visitors does Netflix get every day ?

Data  
Exploration

Data Sources

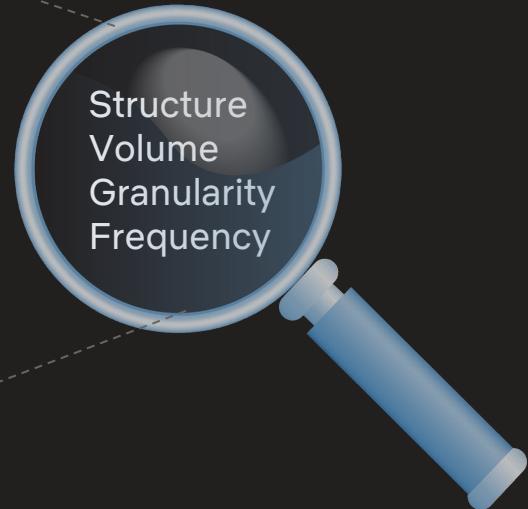


Logs



Data Warehouse

Structure  
Volume  
Granularity  
Frequency



# Problem Statement - How many visitors does Netflix get every day ?

## Data Modeling

### Structure

	country_iso_code	wb_country_code	country_name	country_long_name	region
0	AW	ABW	Aruba	Aruba	Latin America & Caribbean
1	AF	AFG	Afghanistan	Islamic State of Afghanistan	South Asia
2	AO	AGO	Angola	People's Republic of Angola	Sub-Saharan Africa
3	AL	ALB	Albania	Republic of Albania	Europe & Central Asia
4	AD	AND	Andorra	Principality of Andorra	Europe & Central Asia

```
{ "country_iso_code": "Aw",
  "wb_country_code": "ABW",
  "country_name": "Aruba",
  "country_long_name": "Aruba",
  "region": "Latin America And Caribbean"
}
```

### Dimensions

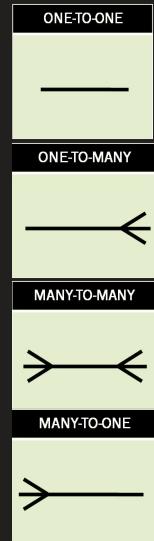


### Metrics

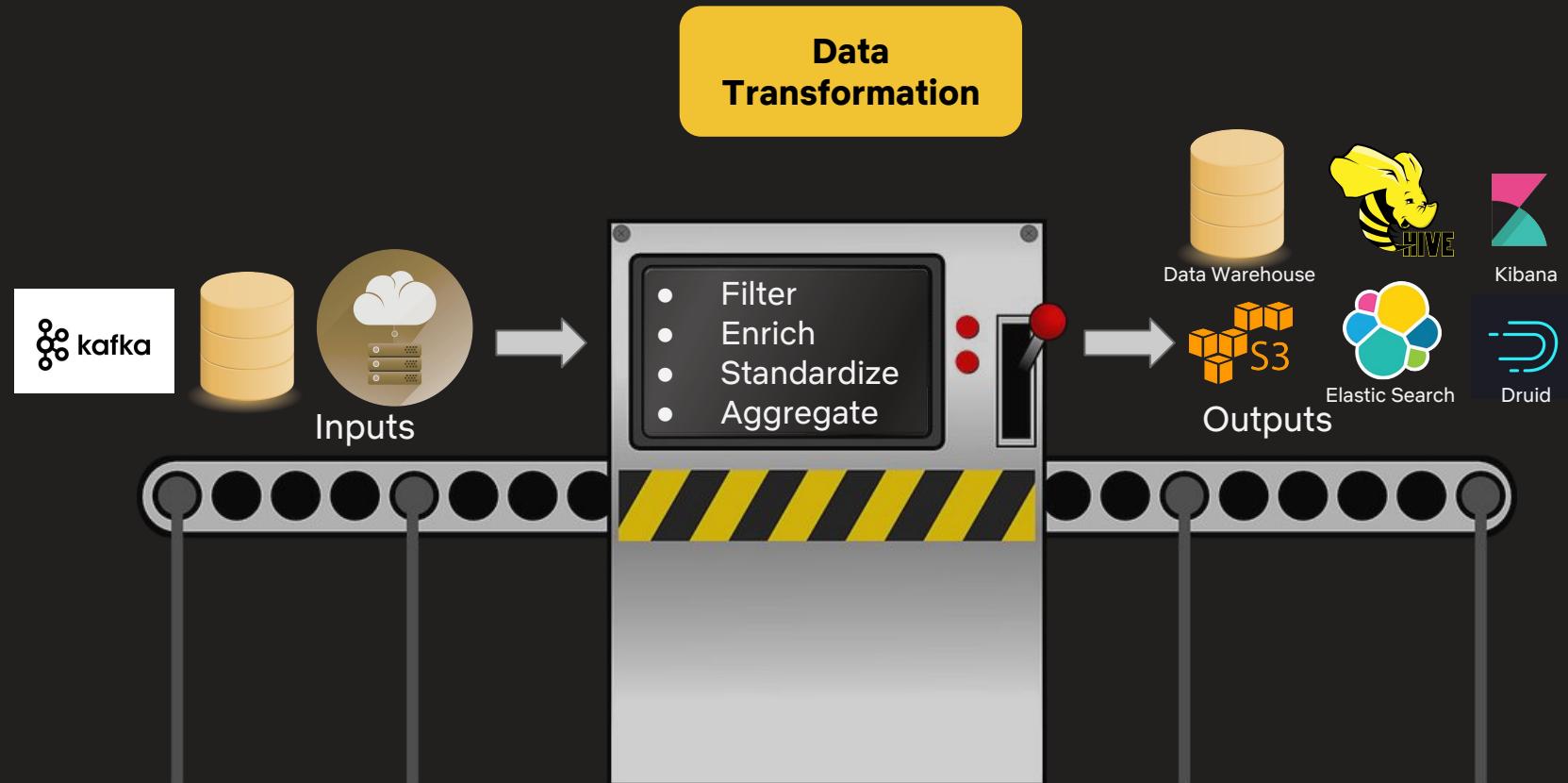


- Visitor Count
- Existing Members
- Repeat Visitors

### Relationships



# Problem Statement - How many visitors does Netflix get every day ?

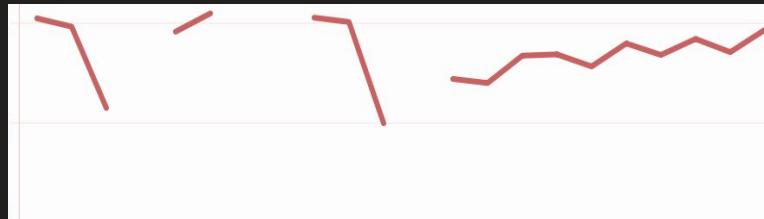


# Problem Statement - How many visitors does Netflix get every day ?

## Data Quality



Trends

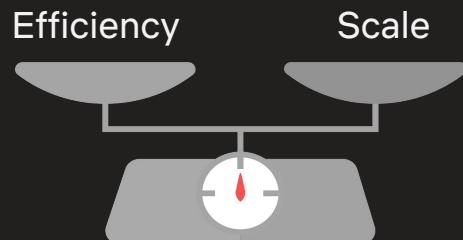
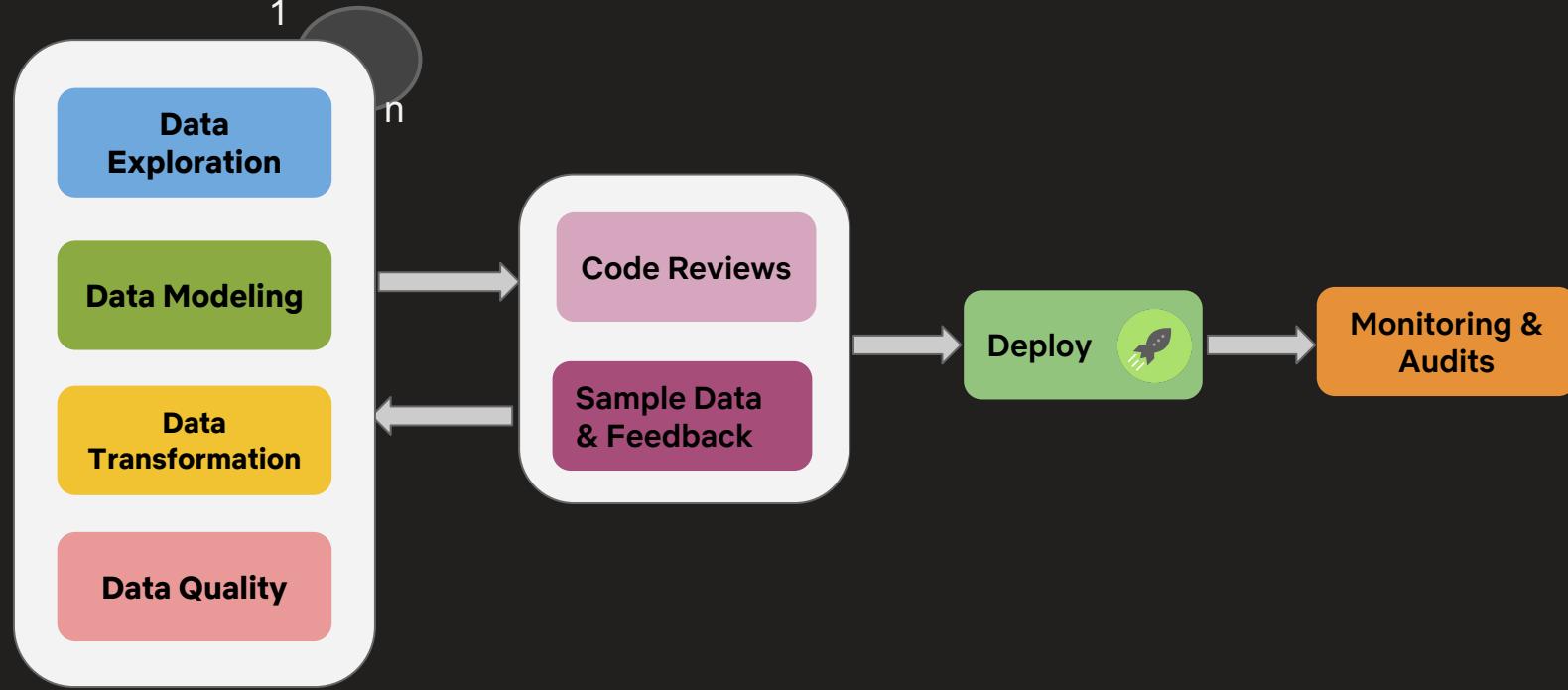


Gaps / Missing Data



Anomalies

# Problem Statement - How many visitors does Netflix get every day ?



# A Typical Data Engineering Project for me

- All the elements we saw before + lot of liaisoning with
  - UI Engineers
  - Data Engineers
  - Data Scientists
  - Data Analysts
  - Data Platform
- Can run from days to several months depending on the project

# A Typical Data Engineering Project for me

- Technologies / Frameworks



Spark



Presto



Druid



Elastic Search



Kibana



Jupyter Notebooks



Scala



Python



Pig



Hive

and more ...

# Questions ?

NETFLIX

# Time to get hands on ...

We'll be using the World Development Indicators Dataset published by World Bank.

<b>Acronym:</b>	WDI
<b>Type:</b>	Time Series
<b>Topics:</b>	Agriculture and Food Security, Climate Change, Economic Growth, Education, Energy and Extractives, Environment and Natural Resources, Financial Sector Development, Gender, Health, Nutrition and Population, Macroeconomic Vulnerability and Debt, Poverty, Private Sector Development, Public Sector Management, Social Development, Social Protection and Labor, Trade, Urban Development
<b>Economy Coverage:</b>	High Income, IBRD, IDA, Low Income, Lower Middle Income, Upper Middle Income
<b>Languages Supported:</b>	English, Arabic, Chinese, French, Spanish
<b>Number of Economies:</b>	217
<b>Geographical Coverage:</b>	World, East Asia & Pacific, American Samoa, Australia, Brunei Darussalam, Cambodia, China <a href="#">More...</a>
<b>External Contact Email:</b>	Data@worldbank.org
<b>Access Options:</b>	Query Tool, API Documentation, Download
<b>Temporal Coverage:</b>	1960 - 2017
<b>Update Frequency:</b>	Quarterly



# Time to get hands on ...

- Spark

Apache Spark is a fast, in-memory data processing engine with elegant and expressive development APIs to allow data workers to efficiently execute streaming, machine learning or SQL workloads that require fast iterative access to datasets.

- Notebooks

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modeling, data visualization, machine learning, and much more.

# Time to get hands on ...

- Go to <https://github.com/annies123/WIBD-Workshops-2018>
- Click on  Clone or download ▾
- Download Zip
- Extract downloaded zip file
- Start the terminal and CD to the location where you extracted the downloaded zip file contents
- Type the following in your terminal to launch notebooks  
jupyter notebook

# Reference Material

- [Spark 2.1.1 Documentation](#)
- [Py Spark Documentation](#)
- [Getting Started With PySpark and Jupyter](#)

Thank you.

NETFLIX