Sparking up Data Engineering

Rohan Sharma
Spark Summit East - Feb 9, 2017

NETFLIX

Agenda

- Context: Netflix
- Netflix Data Ecosystem
- Spark Development @ Netflix
- Stranger Things with Spark
- Q&A

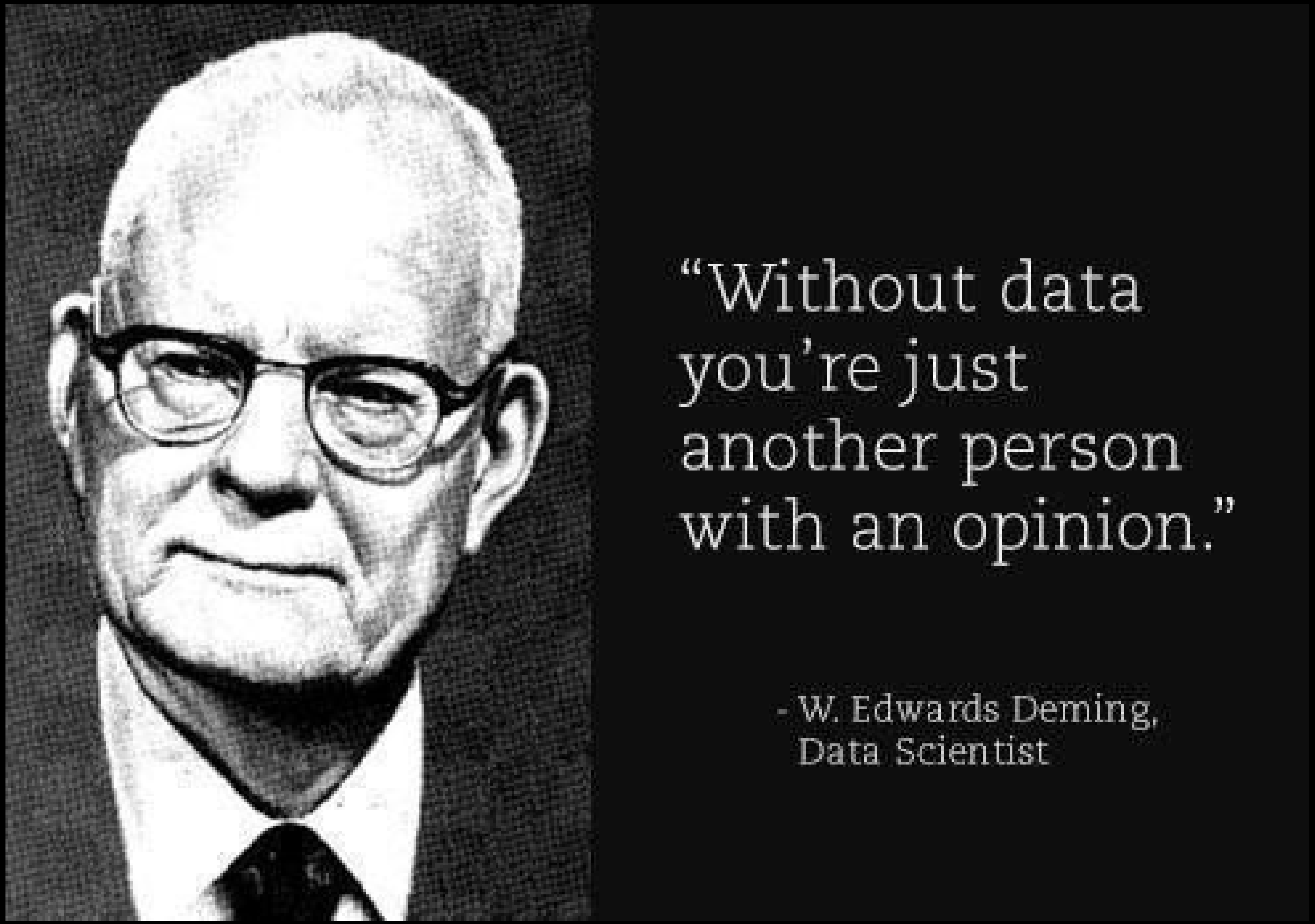
#NetflixEverywhere

- 93+ Million Members
- 190+ Countries
- 125+ Million streaming hours / day
- 1000 hours of Original content in 2017
- ⅓ of US internet traffic during evenings

Netflix Culture

- Freedom and Responsibility
- Context, not Control
- Highly aligned, loosely coupled
- The Paved road

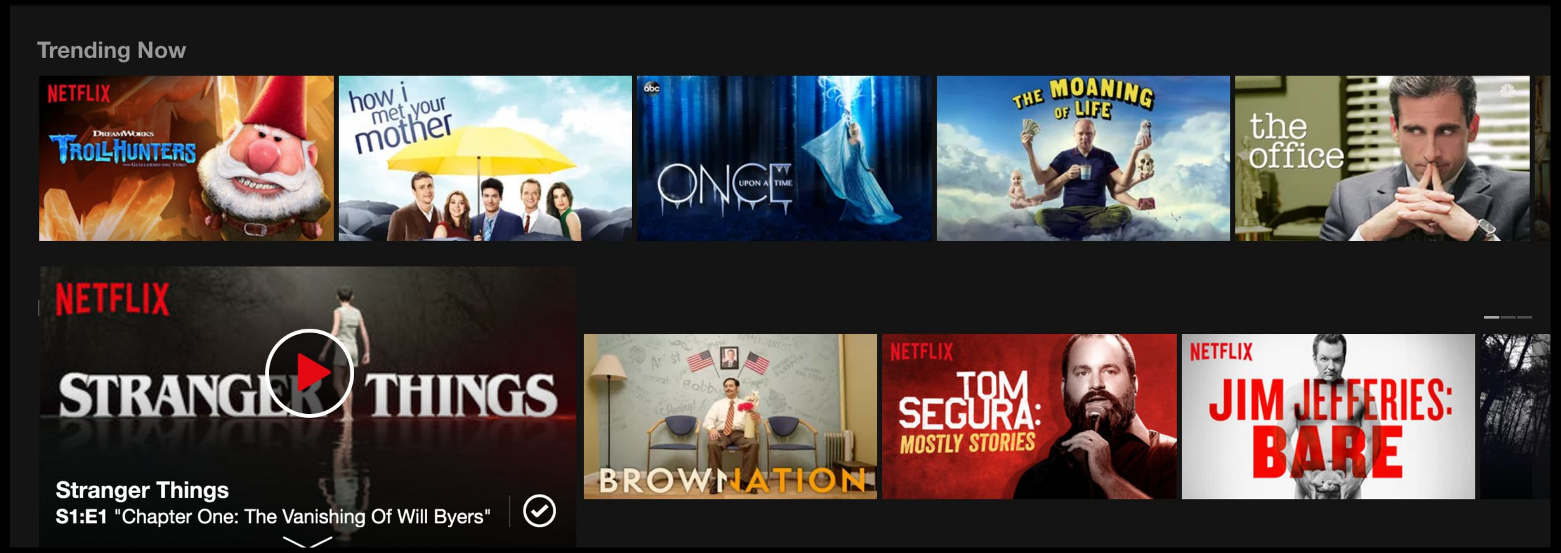
#NetflixData



“Without data you’re just another person with an opinion.”

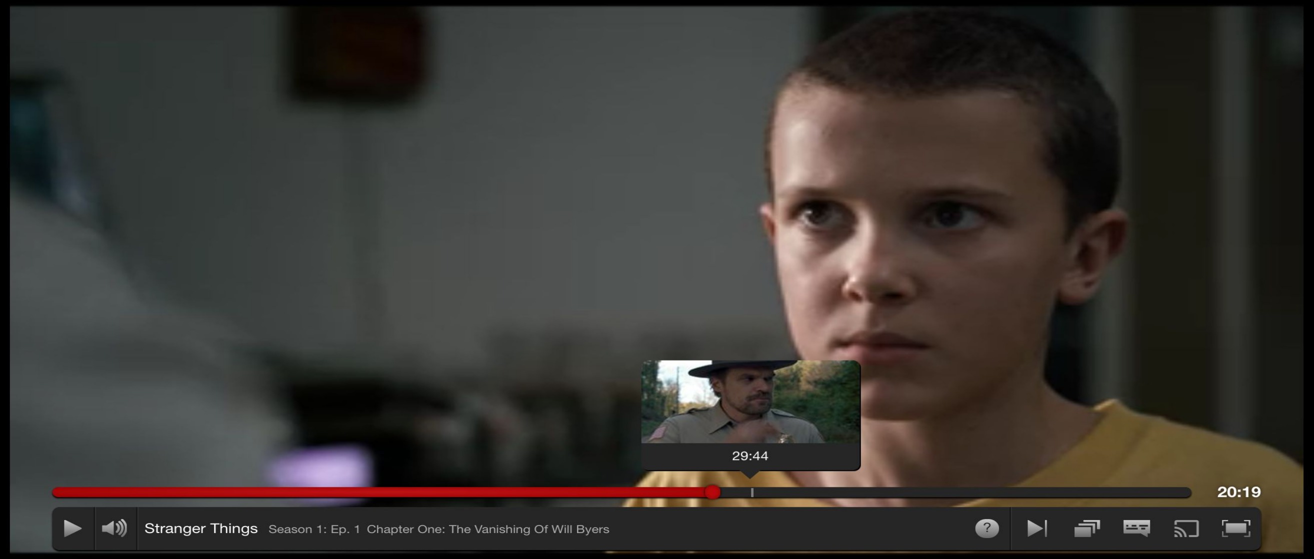
- W. Edwards Deming,
Data Scientist

#NetflixData

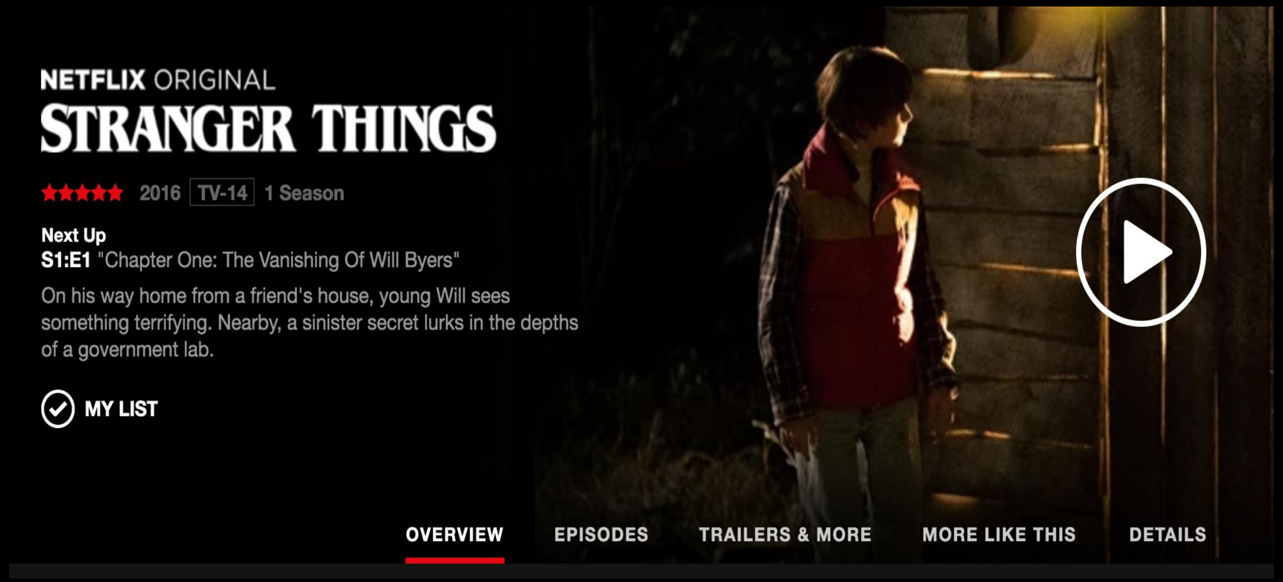

Trending Now

Stranger Things
S1:E1 "Chapter One: The Vanishing Of Will Byers"

Product Experience

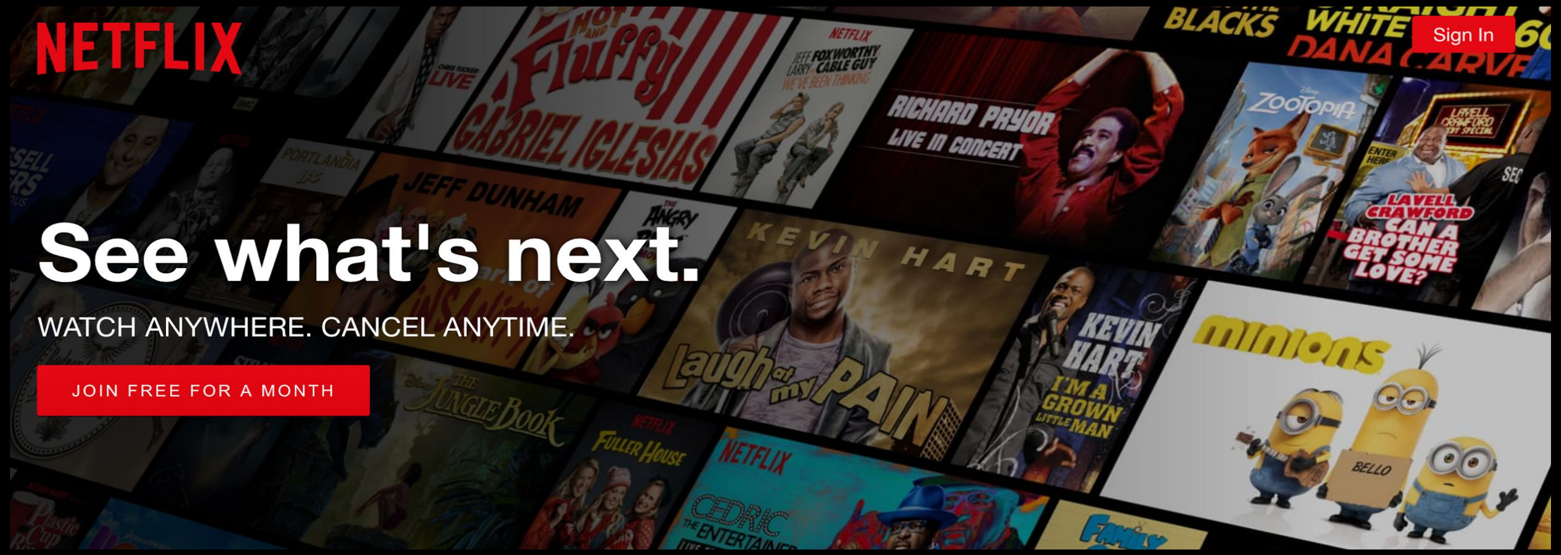

29:44
20:19
Stranger Things Season 1: Ep. 1 Chapter One: The Vanishing Of Will Byers

Streaming Experience


NETFLIX ORIGINAL
STRANGER THINGS
★★★★★ 2016 TV-14 1 Season

Next Up
S1:E1 "Chapter One: The Vanishing Of Will Byers"
On his way home from a friend's house, young Will sees something terrifying. Nearby, a sinister secret lurks in the depths of a government lab.

Content


See what's next.

WATCH ANYWHERE. CANCEL ANYTIME.

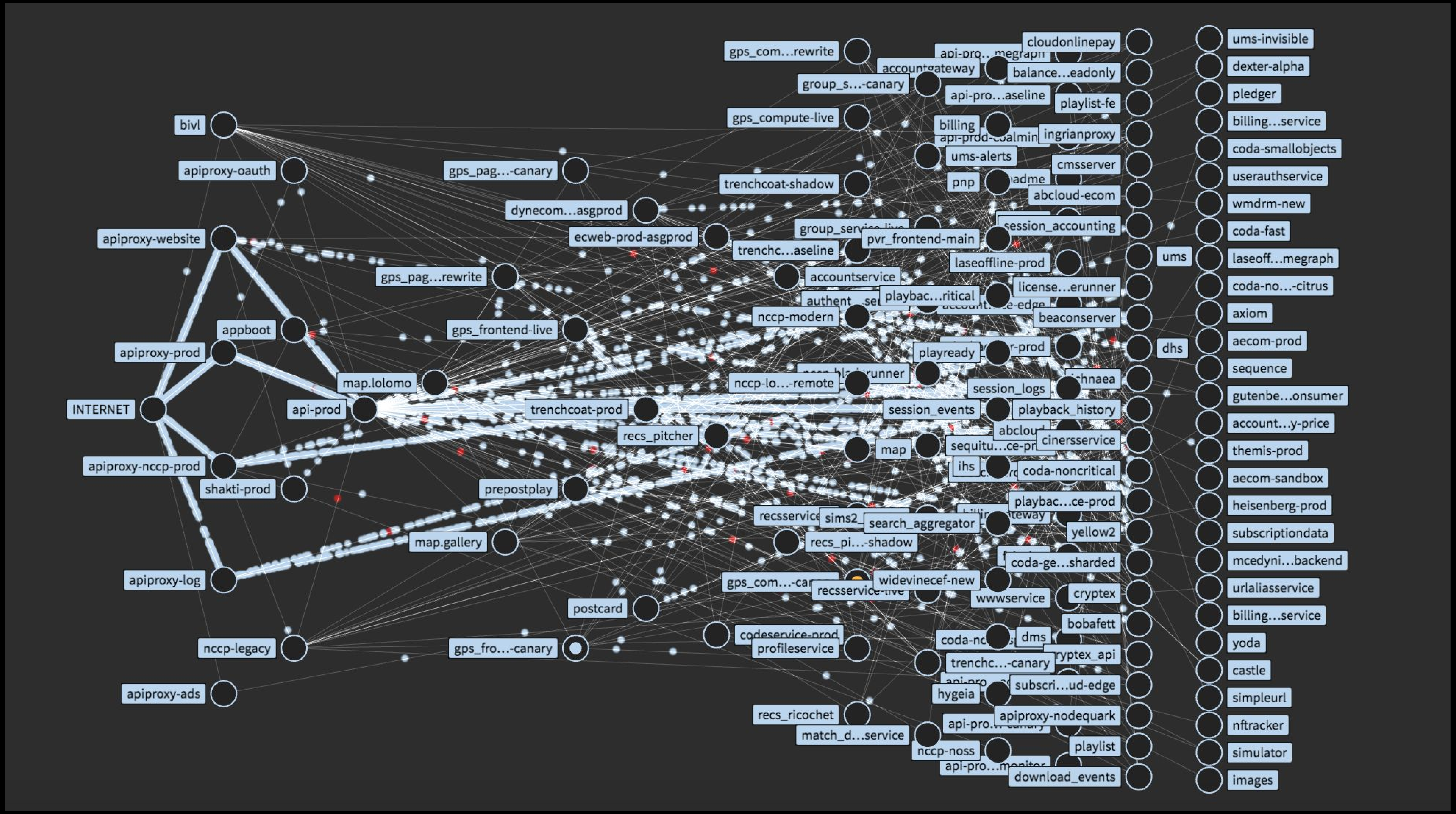
Marketing

Business Operations

Other Functions...

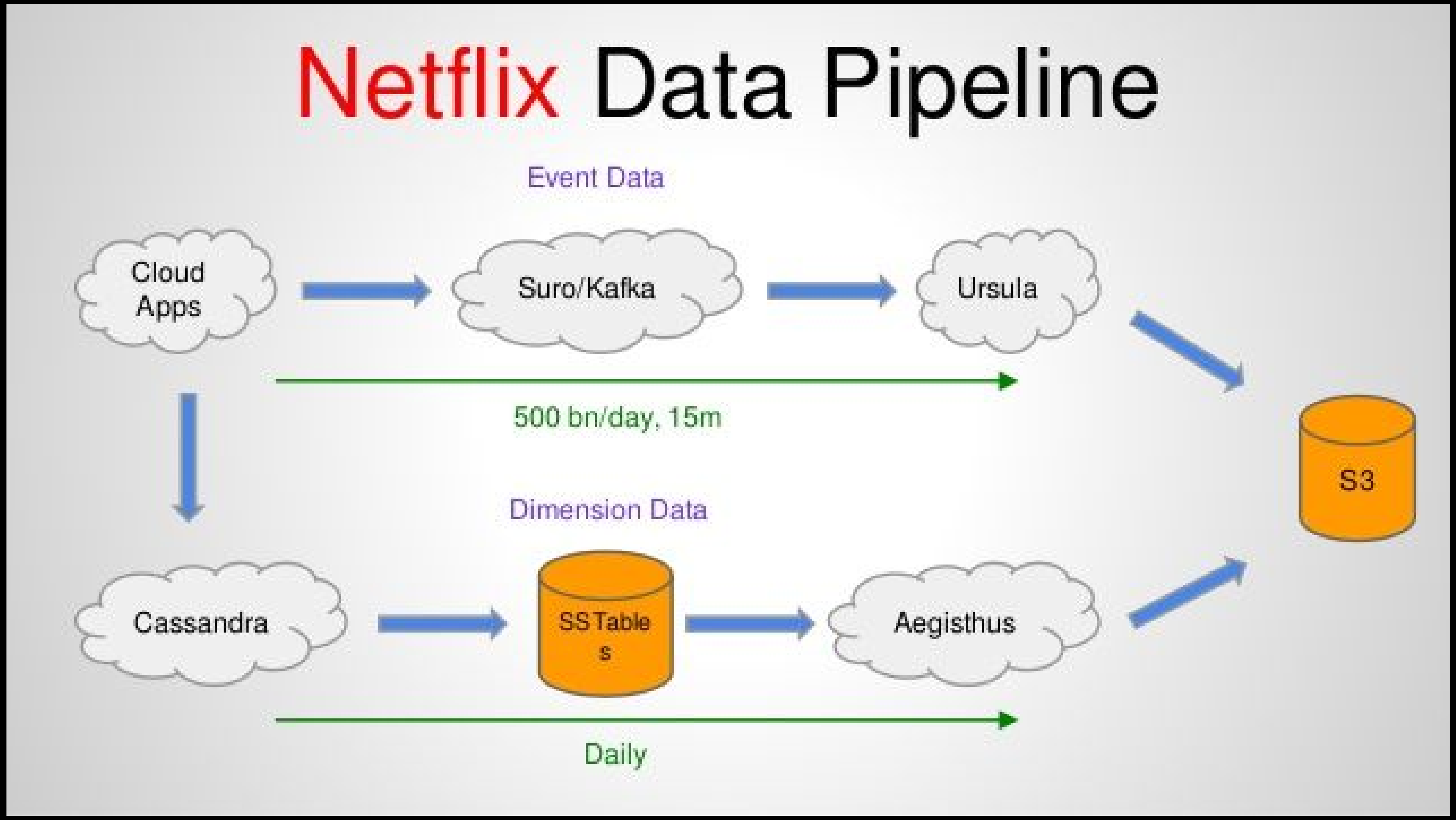
Data Producers

- Member Devices
- CDN Servers
- Application Servers
- Device/Server Telemetry
- Application Data
- Vendors / Partner Data



Data Processing

- Stream Processing - Shriya Arora
- Recommendation Systems - DB Tsai & Gary Yeh
- Batch Processing
- Experimentation Analytics
- Operational Analytics



Data Platform (Batch Processing)

Interface

NETFLIX Big Data Portal

{ … } API

jupyter Zeppelin Notebooks

Tableau

MicroStrategy Micro Strategy

HTML CSS JS JavaScript Applications

Tools

Forklift Transport

Quinto Quality

Sting Visualization

lipstick Pig Workflow Vis

inviso Job/Cluster Vis

NETFLIX OSS

Service

GENIE

Metacat Metadata

UC4 Automic

Compute

Spark

HIVE

presto

druid

Storage

S3

Parquet

druid

TERADATA

Amazon Redshift

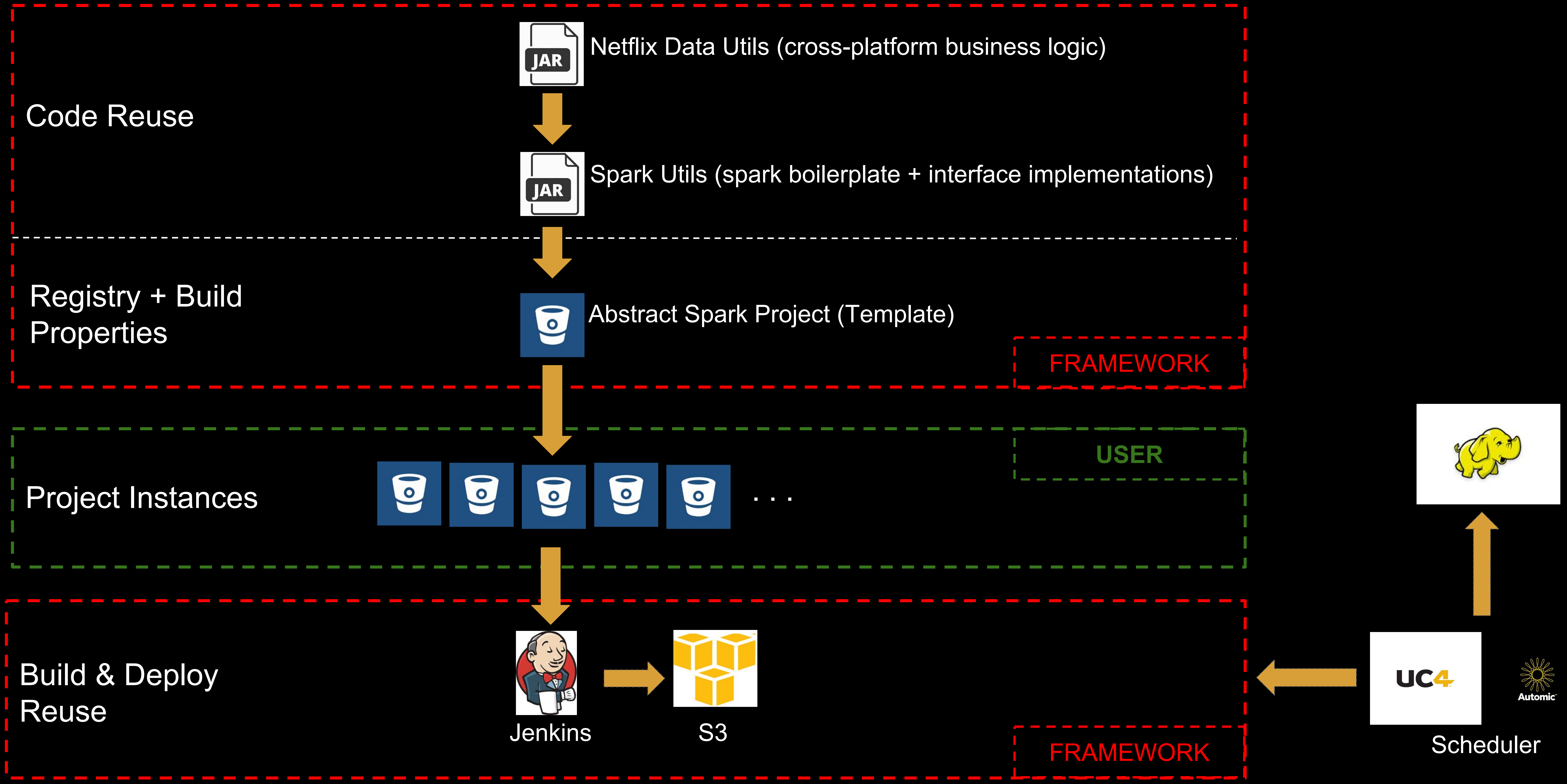
Cloud Warehouse

- 60+ Petabytes
- Hadoop on S3 - separate compute from storage
- Multiple workloads on same cluster
- Tens of thousands of Spark, Pig, Hive, Presto jobs
- Production Cluster : 2600 d2.4xl nodes
- Query Cluster : 400 m4.16xl nodes

Spark Development

- Focus on stakeholders - not process & operations
- Reuse boilerplate code & best practices
- Reuse build & deployment practices
- Make Spark easy-to-use & cruise

Spark Development Template



Code Reuse

Netflix Data Utils (cross-platform business logic)

Spark Utils (spark boilerplate + interface implementations)

Registry + Build Properties

Abstract Spark Project (Template)

FRAMEWORK

USER

Project Instances

. . .

Build & Deploy Reuse

Jenkins

S3

FRAMEWORK

Scheduler

Development Lifecycle

Zeppelin IntelliJ Spark Shell

+

Develop
&
Test

Integration Test

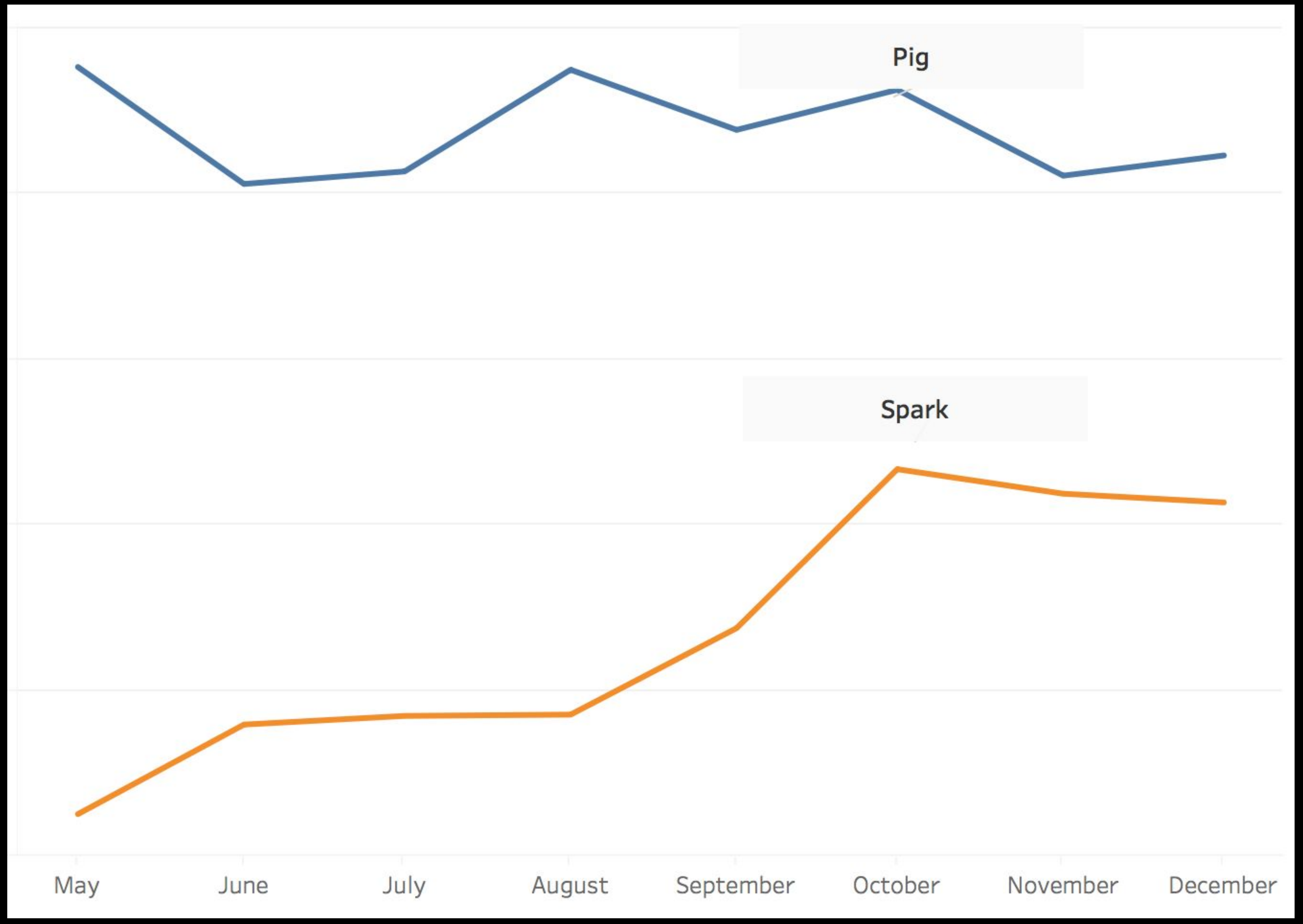
Commit, Build,
Deploy & Run

Bit Bucket Jenkins S3 / Artifactory

Execute

Whats Next...

- Spark added to paved road in early 2016
- Improved query performance / cluster utilization
- Pyspark templates & data utils



STRANGER THINGS

Code Refactoring

- Download Source Code
- Semi-Compile to Abstract Syntax Trees
- Check for re-factoring rules
- Codegen refactored lines and update source

- Invoke dockerized CI service to build
- Raise Pull Requests with inferred reviewers

More Details:
Jonathan Schneider

Linked In:
https://www.linkedin.com/in/jonkschneider/

Youtube:
https://www.youtube.com/watch?v=JbcKFKiBU60

Questions

To learn more, please visit:

youtube channel
NetflixData

twitter handle
@NetflixData