Building Streaming Data Pipelines

Using Azure Cloud Services

Rolf Tesmer
Data Solution Architect (DSA)
Melbourne, Australia

Linked In: https://www.linkedin.com/in/rolftesmer/
Blog: https://mrfoxsql.wordpress.com/

Agenda

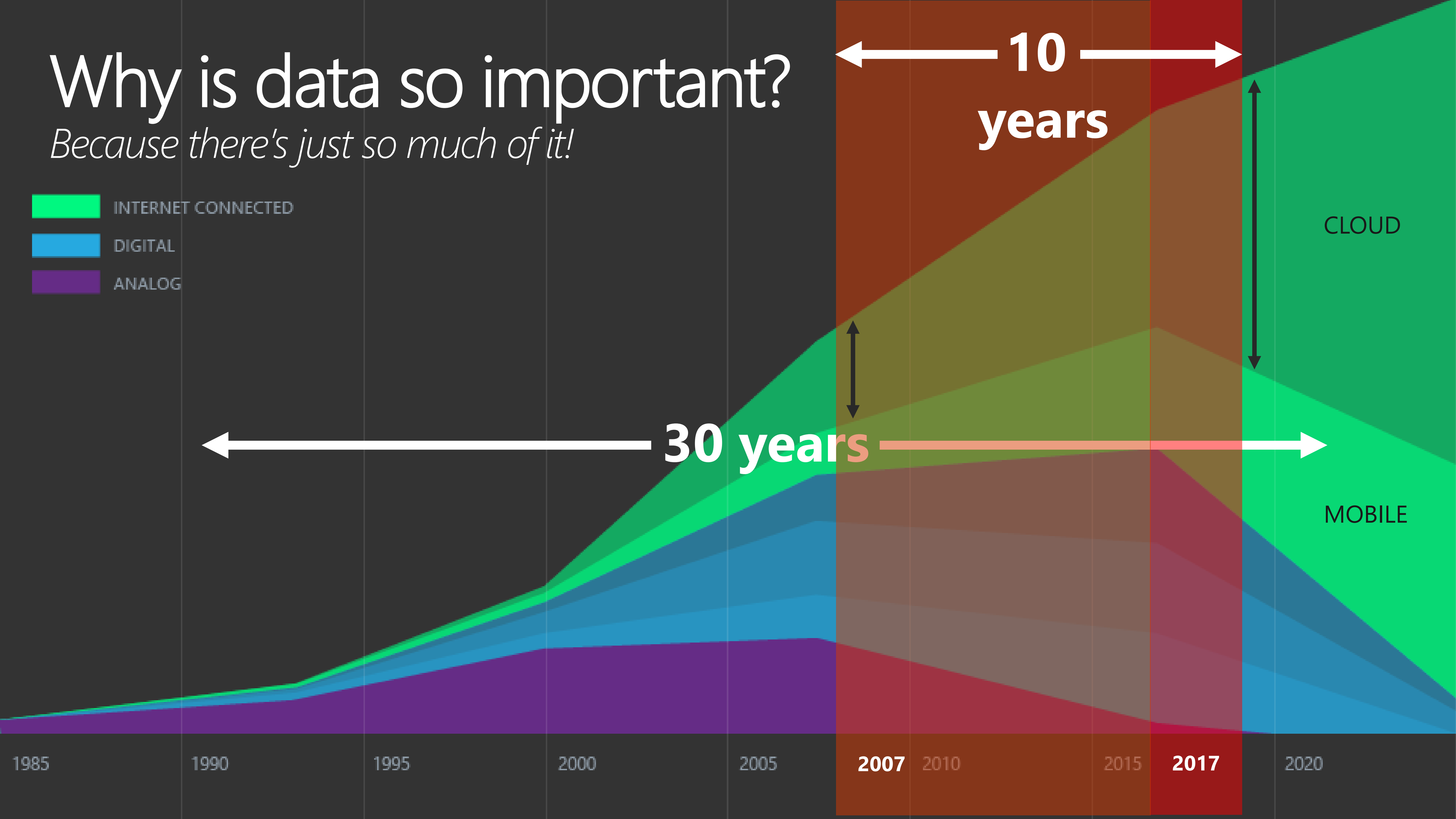
- Introduction
- What exactly is the data platform nowadays?
- Data pipeline services and options in Azure
- **Demonstration: *Lets see a data pipeline!***
- What’s next? Wrap up and summary

Agenda

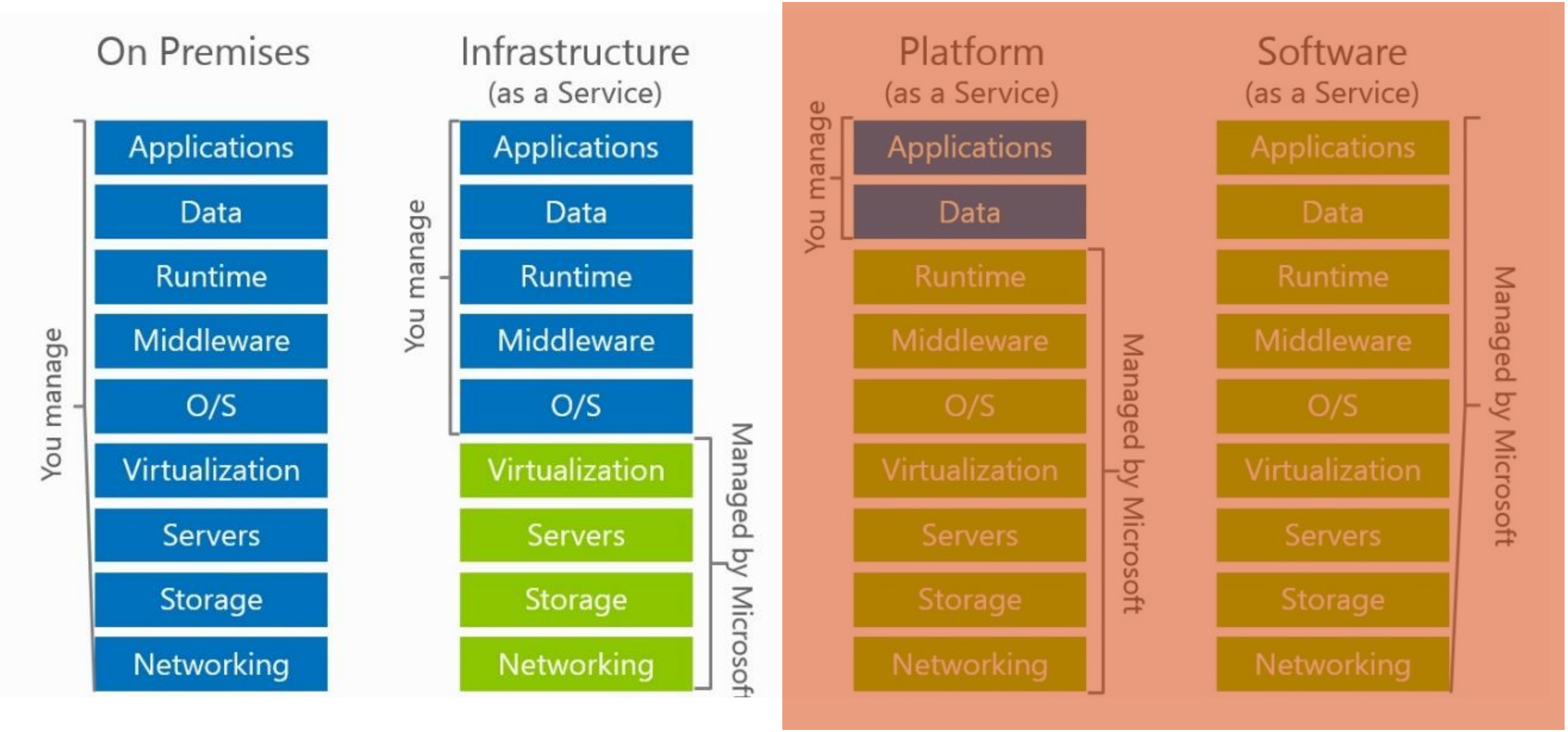
- Introduction

Why is data so important?

Because there's just so much of it!



On-Prem *vs* IaaS *vs* PaaS *vs* SaaS – *Which One?*



On Premises

You manage

- Applications
- Data
- Runtime
- Middleware
- O/S
- Virtualization
- Servers
- Storage
- Networking

Infrastructure
(as a Service)

You manage

- Applications
- Data
- Runtime
- Middleware
- O/S

Managed by Microsoft

- Virtualization
- Servers
- Storage
- Networking

Platform
(as a Service)

You manage

- Applications
- Data

Managed by Microsoft

- Runtime
- Middleware
- O/S
- Virtualization
- Servers
- Storage
- Networking

Software
(as a Service)

Managed by Microsoft

- Applications
- Data
- Runtime
- Middleware
- O/S
- Virtualization
- Servers
- Storage
- Networking

http://azureplatform.azurewebsites.net/

* Preview Services

Compute

- Virtual Machines
- Virtual Machine Scale Sets
- Azure Container Service
- Azure Container Registry
- Functions
- Batch
- Service Fabric
- Cloud Services

Networking

- Virtual Network
- Load Balancer
- Application Gateway
- VPN Gateway
- Azure DNS
- Traffic Manager
- ExpressRoute
- Network Watcher

Storage

- Storage: Blobs, Tables, Queues, Files, Disks
- Data Lake Store
- StorSimple
- Azure Backup
- Site Recovery

Web & Mobile

- Web Apps
- Mobile Apps
- Logic Apps
- API Apps
- Content Delivery Network
- Media Services
- Search

Databases

- SQL Database
- SQL Data Warehouse
- SQL Server Stretch Database
- DocumentDB
- Redis Cache
- Data Factory

Intelligence & Analytics

- HDInsight
- Machine Learning
- Cognitive Services
- Azure Bot Service*
- Data Lake Analytics
- Power BI Embedded
- Azure Analysis Services

Internet of Things & Enterprise Integration

- Azure IoT Hub
- Event Hubs
- Stream Analytics
- Notification Hubs
- BizTalk Services
- Service Bus
- Data Catalog

Security + Identity

- Security Center
- Key Vault
- Azure Active Directory
- B2C
- Domain Services
- Multi-Factor Authentication

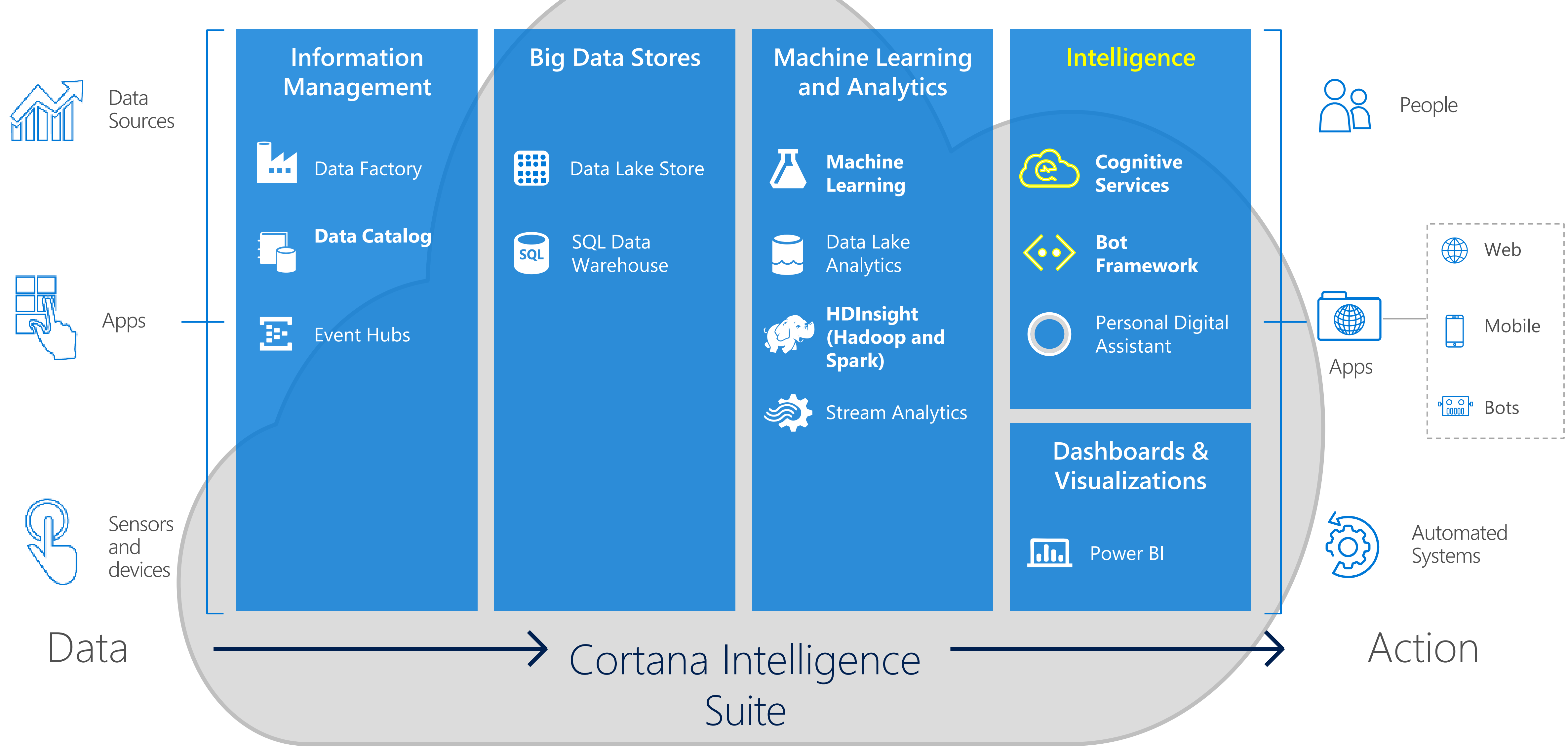
Developer Services

- Visual Studio Team Services
- Azure DevTest Labs
- VS Application Insights
- API Management
- HockeyApp
- Developer Tools
- Service Profiler*

Monitoring & Management

- Azure Portal
- Azure Resource Manager
- Azure Advisor
- Azure Monitor
- Log Analytics
- Automation
- Scheduler

Cortana Intelligence Suite



Data Sources

Apps

Sensors and devices

Data

Information Management

Data Factory

Data Catalog

Event Hubs

Big Data Stores

Data Lake Store

SQL Data Warehouse

Machine Learning and Analytics

Machine Learning

Data Lake Analytics

HDInsight (Hadoop and Spark)

Stream Analytics

Intelligence

Cognitive Services

Bot Framework

Personal Digital Assistant

Dashboards & Visualizations

Power BI

Cortana Intelligence Suite

People

Apps

Web

Mobile

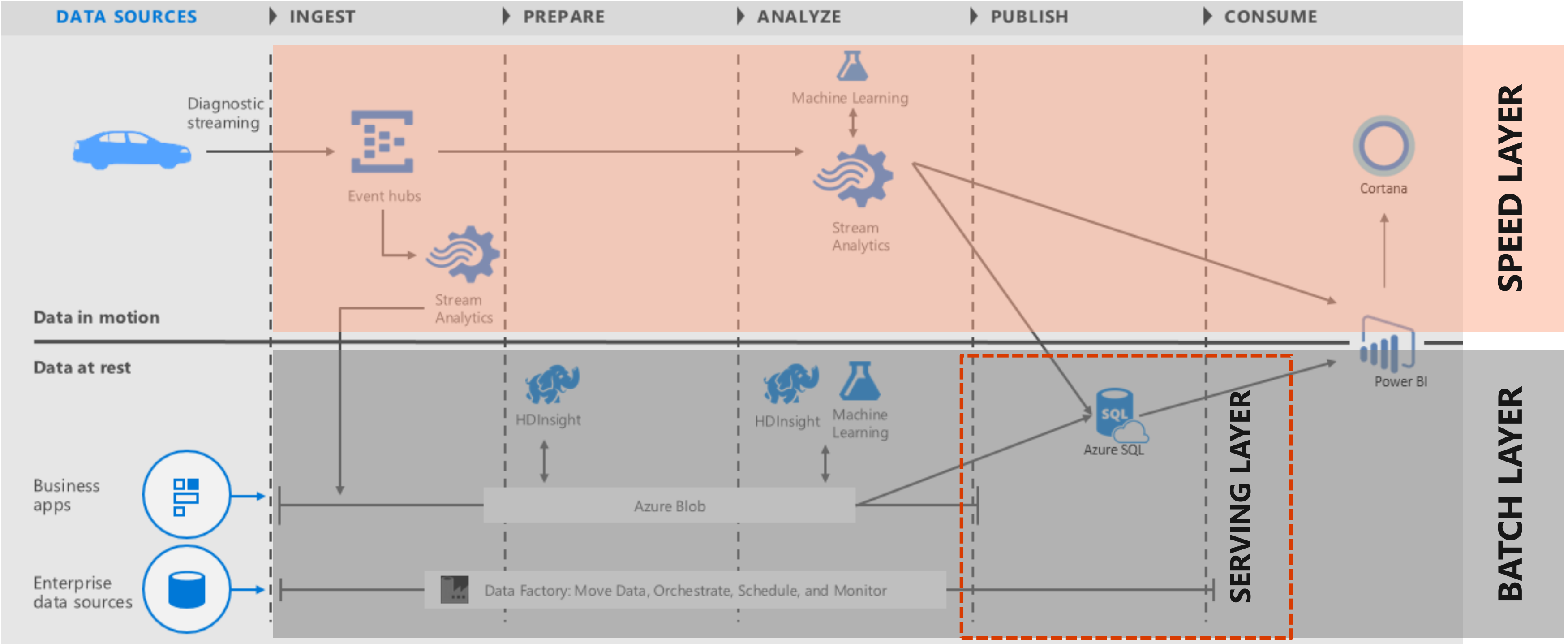
Bots

Automated Systems

Action

What is the LAMBDA architecture?

"The Objective of ***Lambda Architecture*** *is to leverage the combined power of both* ***batch*** *&* ***real-time*** *processing to address the business scenarios where it requires both* ***historic view of the data*** *as well as getting insight into the* ***data in real-time*** *as business happens."*



https://social.technet.microsoft.com/wiki/contents/articles/33626.lambda-architecture-implementation-using-microsoft-azure.aspx
https://gallery.cortanaintelligence.com/Solution/Telemetry-Analytics
https://docs.microsoft.com/en-us/azure/machine-learning/cortana-analytics-playbook-vehicle-telemetry

What exactly is a “data pipeline” anyway?

- Different definitions **depending on which vendor you talk to**
- **Microsoft** have **no formal definition**
- *But*… a couple of definitions that **I like**…

*“**pipelines** are formed from multiple individual ‘fit for purpose’ services aligned in sequences that perform a set of specific targeted actions on data that is typically in transit.”*

Source: (Rolf Tesmer) ☺

*"a **pipeline** is a set of data processing elements connected in series, where the output of one element is the input of the next one. The elements of a **pipeline** are often executed in parallel or in time-sliced fashion"*

Source: (Wikipedia)

*"a data **pipeline** is the software that consolidates data from multiple sources and makes it available to be used strategically"*

Source: (Unknown Original Source)

Where did this come from, and why do we care?

1. Customers are on a **multi-year transformational journey**

2. Many **data sources** are **not static** or **at rest**

3. Solutions **cannot wait** for data to be landed before using it

4. building pipelines…

- **Historically** → Complex, costly, time consuming
- **Today** → Fast, simple, “fit for purpose” services from same **data platform**

As modern day Data Professionals we have to deal with it

Agenda

- What is considered the new data platform

What *was* the data platform?

Up till **~5-10 years ago** it was a **central relational platform**

…*and*… **included relational-like services** (OLTP, OLAP, DW, ETL, MDM, +)

…*and*… often **on-prem**, or in a hosted DC

…*and*… rarely hosted in external **public cloud** providers (Azure, AWS, +)

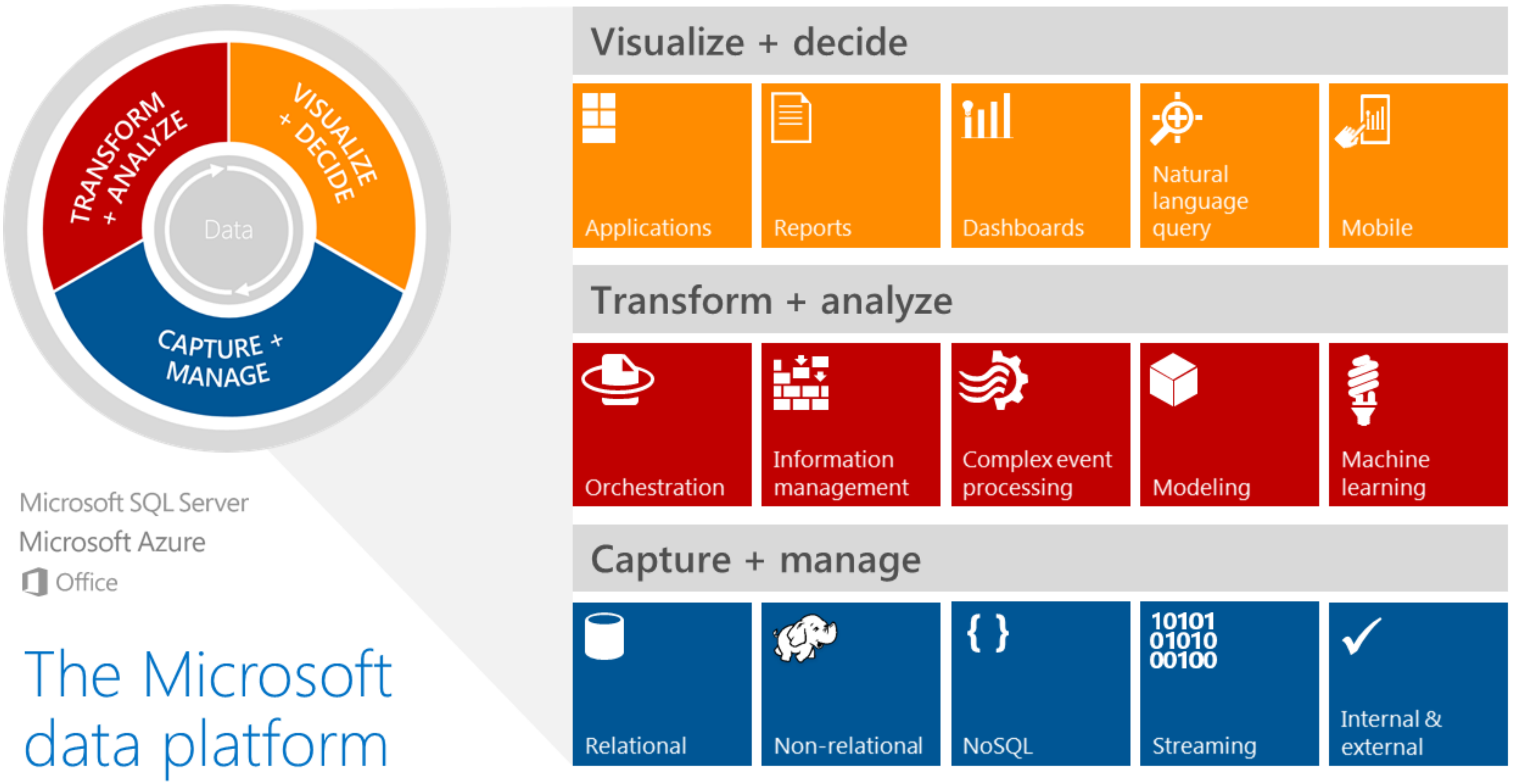
Occasionally included **special projects** (ie Big Data, NoSQL, IoT)

https://mrfoxsql.wordpress.com/2017/04/19/what-exactly-is-the-data-platform-nowadays/

What *is* the data platform *now*?

- *Mix of…* **on-prem and public cloud**
- *Mix of…* **deployment models (IaaS, PaaS, SaaS)**
- *Mix of…* **specific “fit for purpose” individual data services**

- These services are across a range of uses including;
  1. Ingestion
  2. Transformation
  3. Storage
  4. Analytics
  5. Visualisation



Visualize + decide

Applications | Reports | Dashboards | Natural language query | Mobile

Transform + analyze

Orchestration | Information management | Complex event processing | Modeling | Machine learning

Capture + manage

Relational | Non-relational | NoSQL | Streaming | Internal & external

Microsoft SQL Server
Microsoft Azure
Office

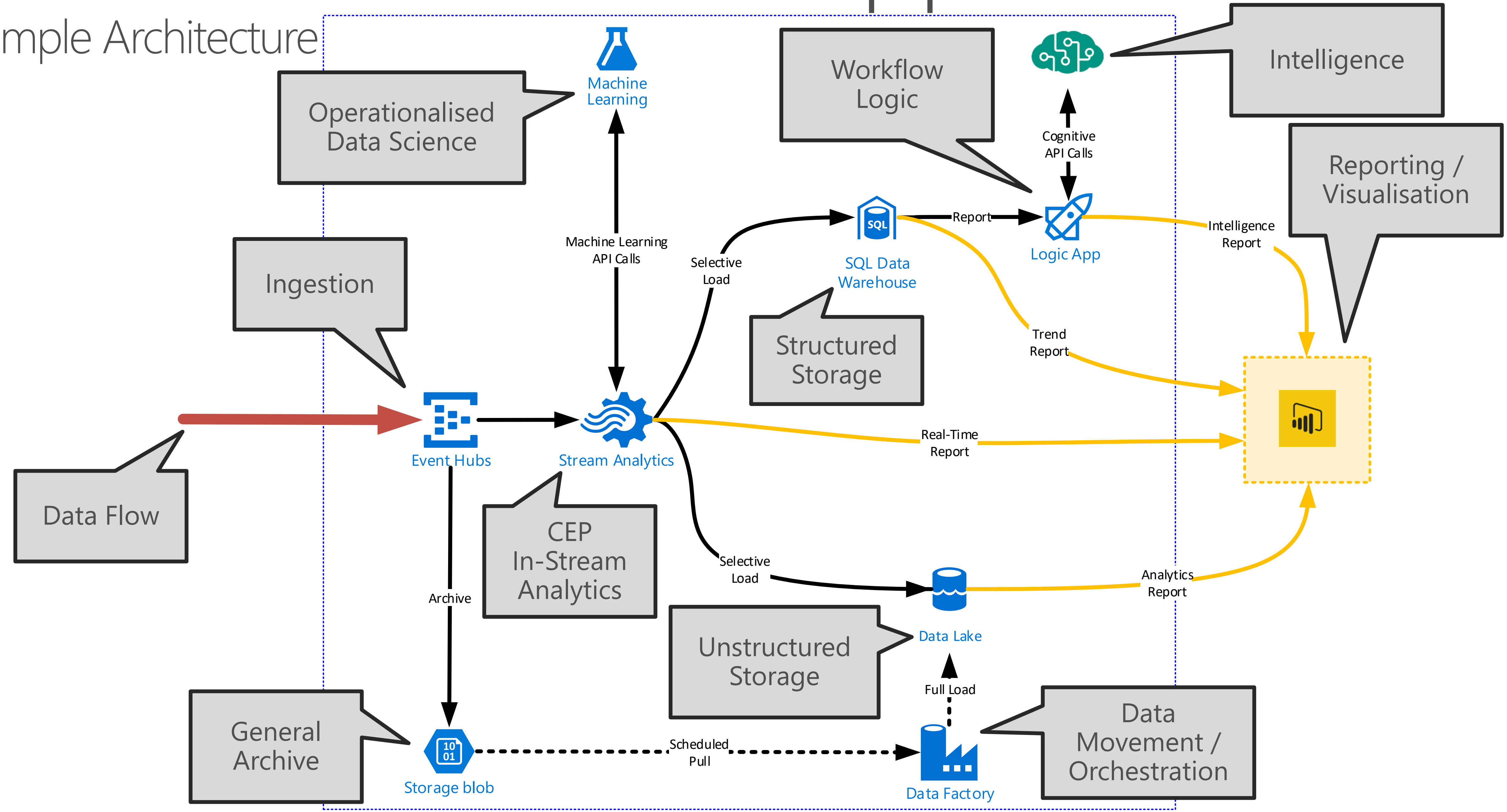
The Microsoft data platform

Agenda

- Data pipeline services and options in Azure

What are some of the Azure pipeline services?

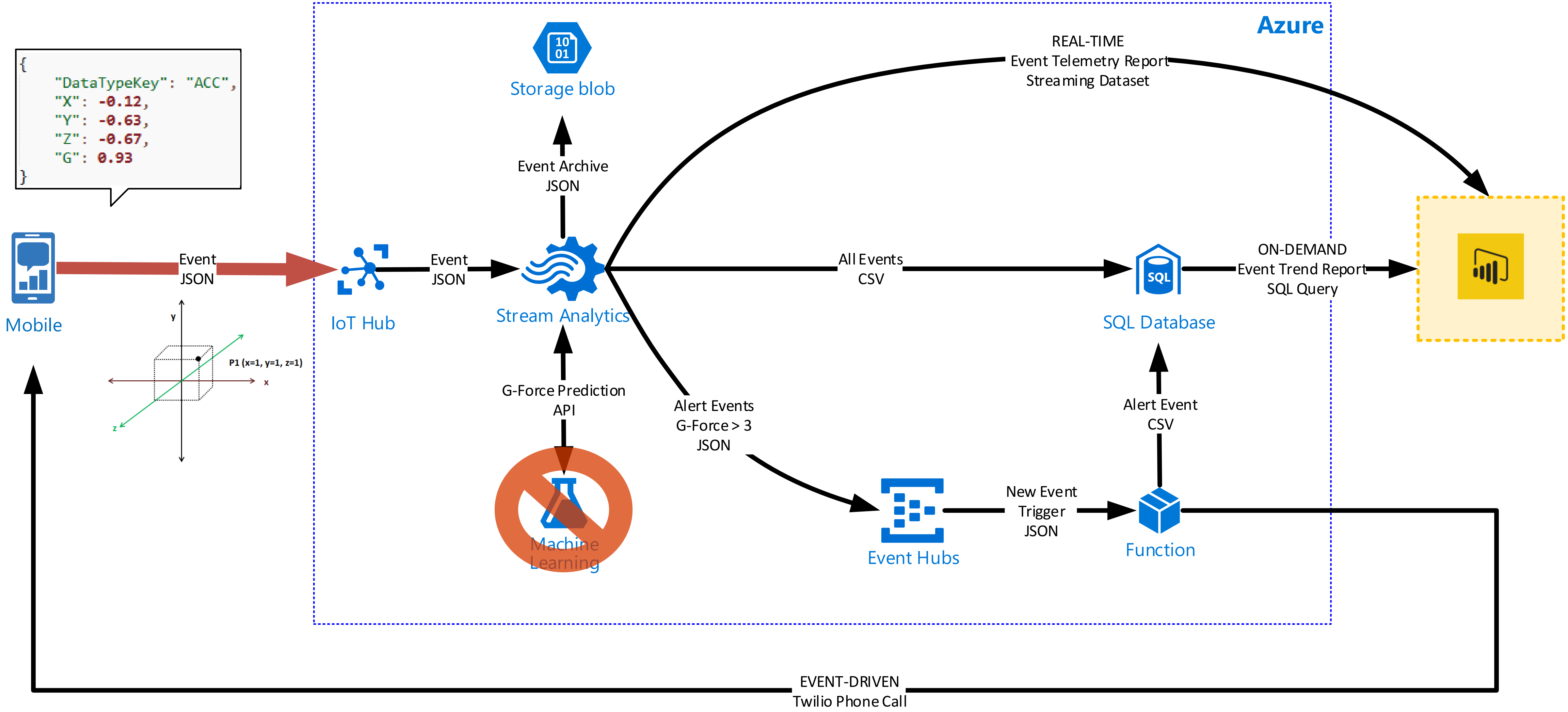
Example Architecture



Agenda

- Demos / Examples: Lets see some Azure pipelines!

Demonstration → Mobile G-Force Solution - !



```
{
    "DataTypeKey": "ACC",
    "X": -0.12,
    "Y": -0.63,
    "Z": -0.67,
    "G": 0.93
}
```

Azure

Mobile

Event
JSON

IoT Hub

Event
JSON

Stream Analytics

Storage blob

Event Archive
JSON

G-Force Prediction
API

Machine
Learning

REAL-TIME
Event Telemetry Report
Streaming Dataset

All Events
CSV

SQL Database

ON-DEMAND
Event Trend Report
SQL Query

Alert Events
G-Force > 3
JSON

Event Hubs

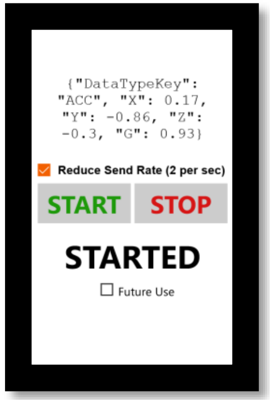
New Event
Trigger
JSON

Function

Alert Event
CSV

EVENT-DRIVEN
Twilio Phone Call

Demonstration → Mobile G-Force Solution - !

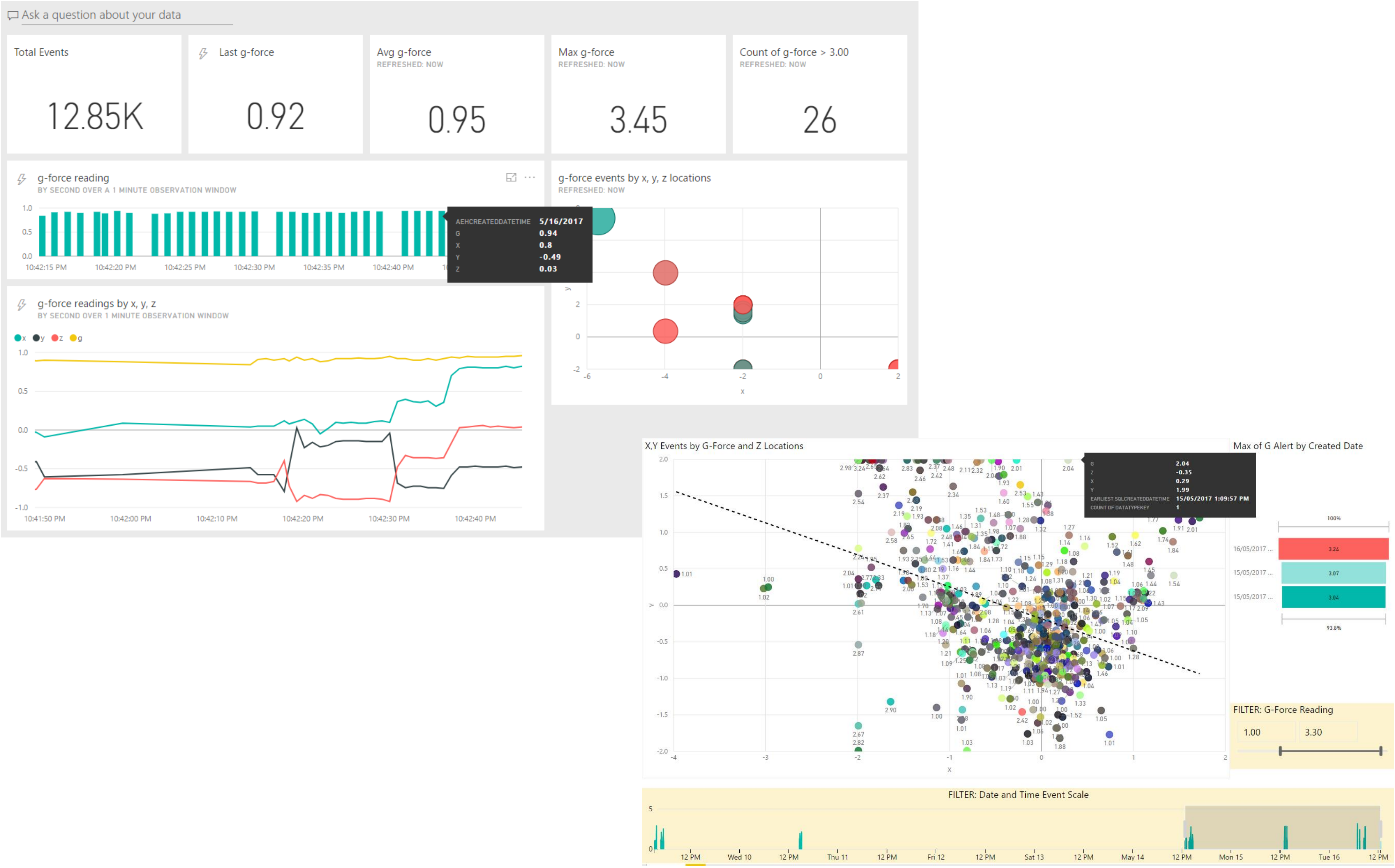


```
{"DataTypeKey":
"ACC", "X": 0.17,
"Y": -0.86, "Z":
-0.3, "G": 0.93}
```

☑ Reduce Send Rate (2 per sec)

STARTED

□ Future Use



Ask a question about your data

Total Events 12.85K

Last g-force 0.92

Avg g-force REFRESHED: NOW 0.95

Max g-force REFRESHED: NOW 3.45

Count of g-force > 3.00 REFRESHED: NOW 26

g-force reading BY SECOND OVER A 1 MINUTE OBSERVATION WINDOW

AEHCREATEDDATETIME	5/16/2017
G	0.94
X	0.8
Y	-0.49
Z	0.03

g-force events by x, y, z locations REFRESHED: NOW

g-force readings by x, y, z BY SECOND OVER 1 MINUTE OBSERVATION WINDOW

X,Y Events by G-Force and Z Locations

G	2.04
Z	-0.35
X	0.29
Y	1.99
EARLIEST SQLCREATEDDATETIME	15/05/2017 1:09:57 PM
COUNT OF DATATYPEKEY	1

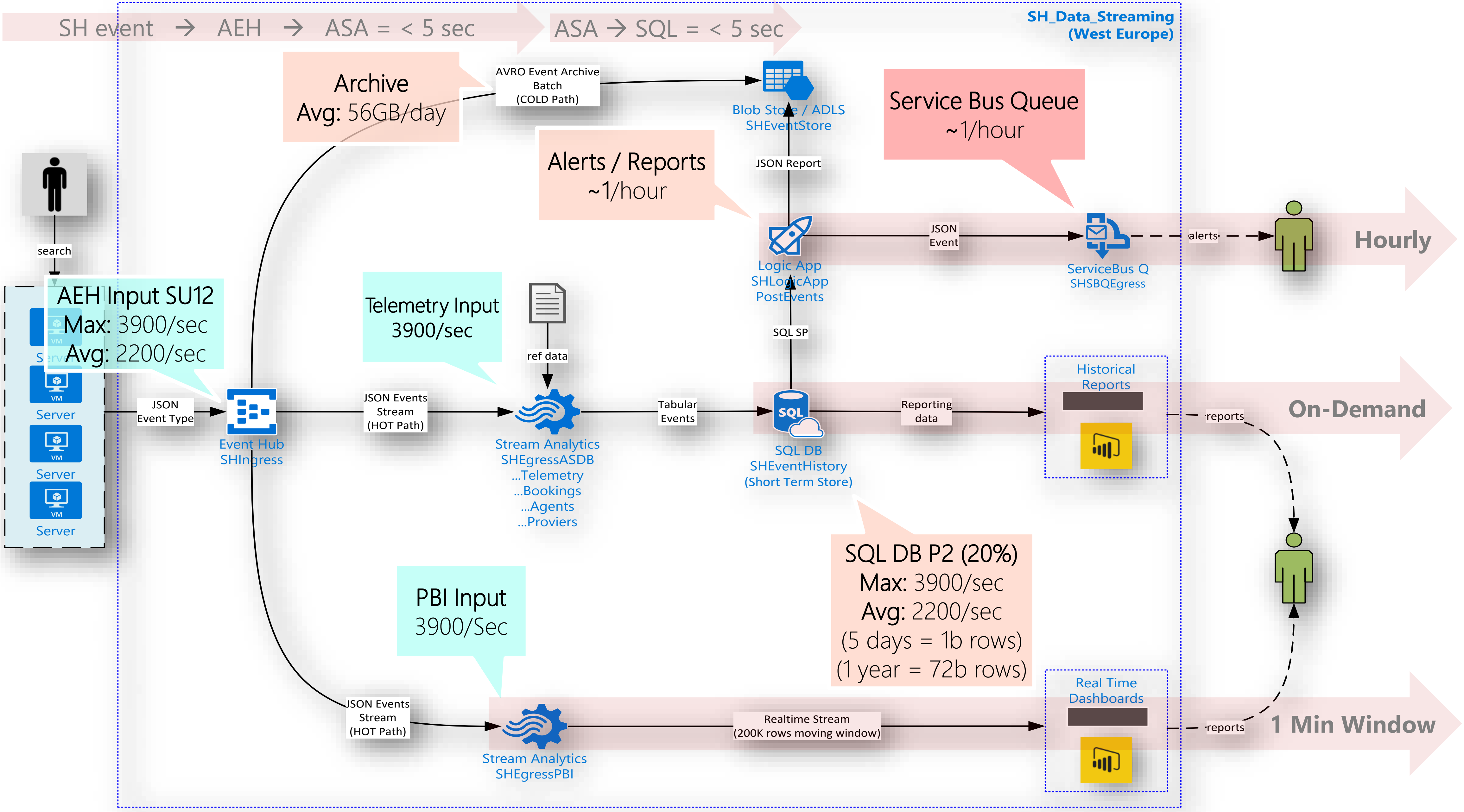
Max of G Alert by Created Date

16/05/2017 ...	3.24
15/05/2017 ...	3.07
15/05/2017 ...	3.04

FILTER: G-Force Reading 1.00 3.30

FILTER: Date and Time Event Scale

Other Examples → High Scale Web Search Telemetry



SH event → AEH → ASA = < 5 sec

ASA → SQL = < 5 sec

Archive
Avg: 56GB/day

Alerts / Reports
~1/hour

Service Bus Queue
~1/hour

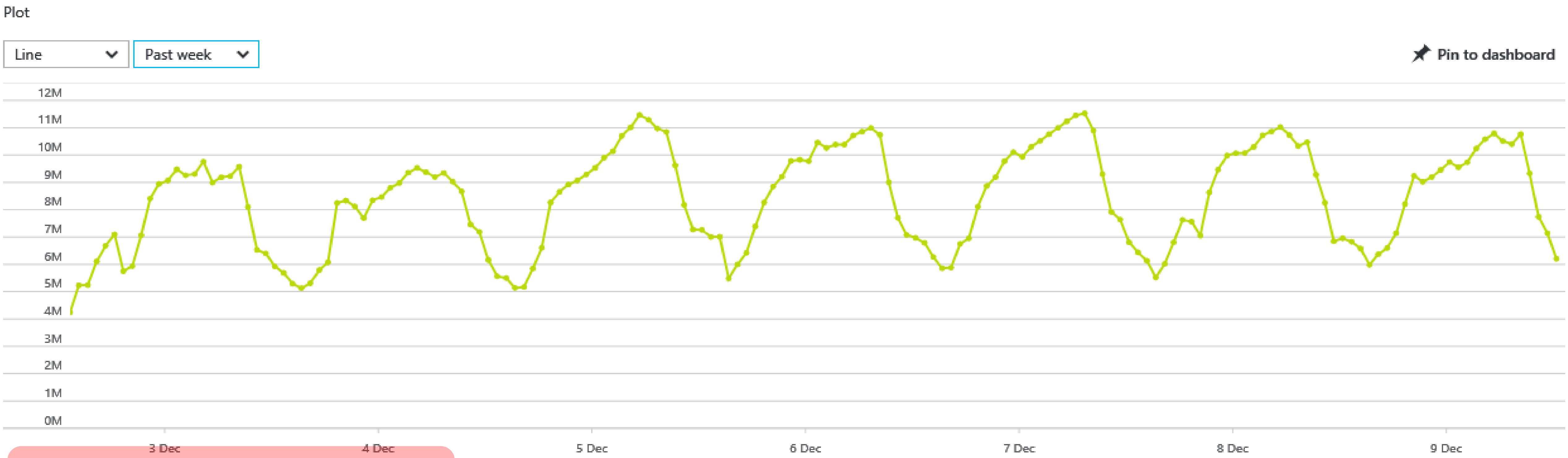
AEH Input SU12
Max: 3900/sec
Avg: 2200/sec

Telemetry Input
3900/sec

PBI Input
3900/Sec

SQL DB P2 (20%)
Max: 3900/sec
Avg: 2200/sec
(5 days = 1b rows)
(1 year = 72b rows)

Web Search Telemetry – Inbound Messages



Plot

Line

Past week

INCOMING MESSAGES
1.41 B

Average Load → 1,410,000,000 / week

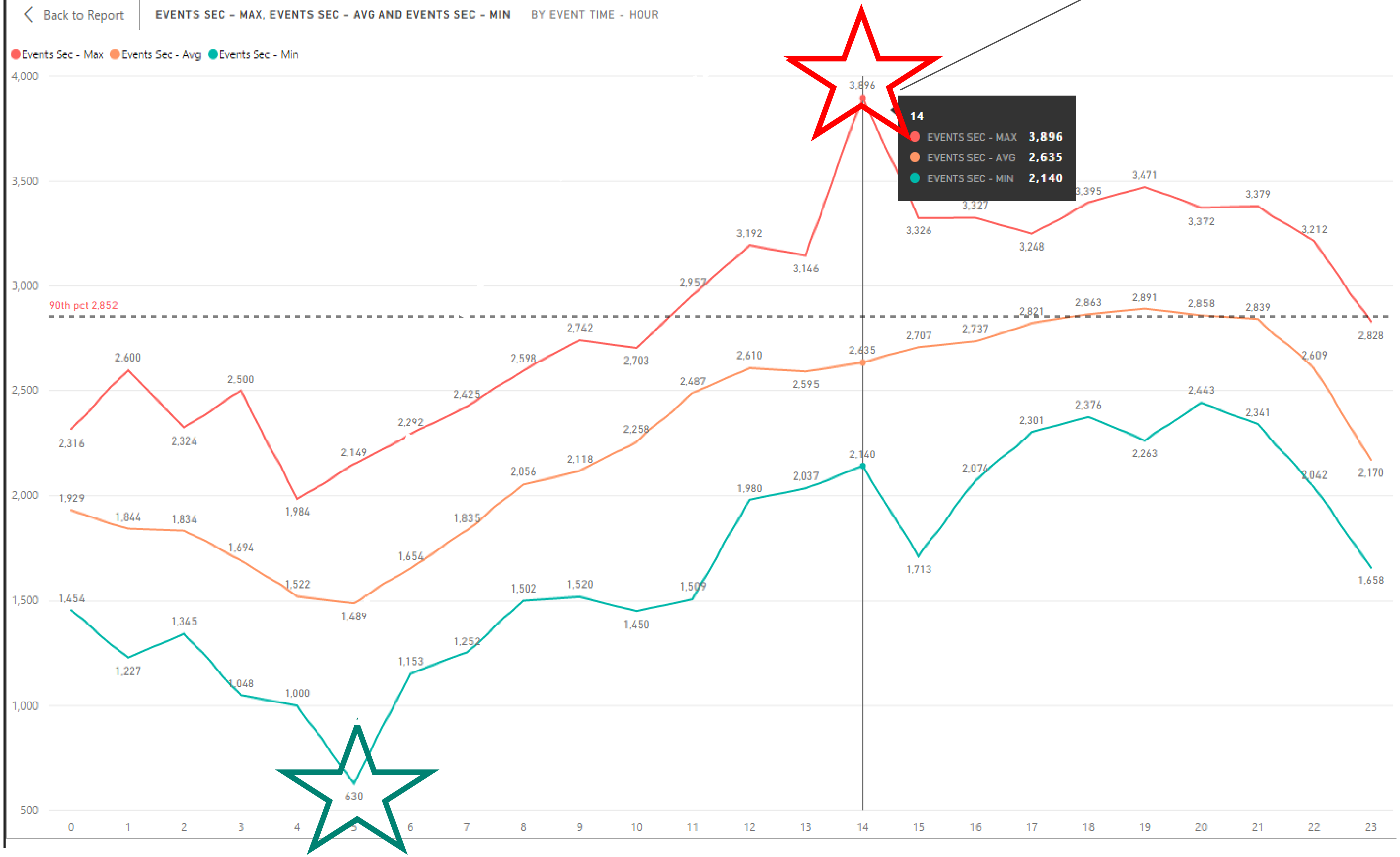
= 201,000,000 / day

= 8,392,000 / hour

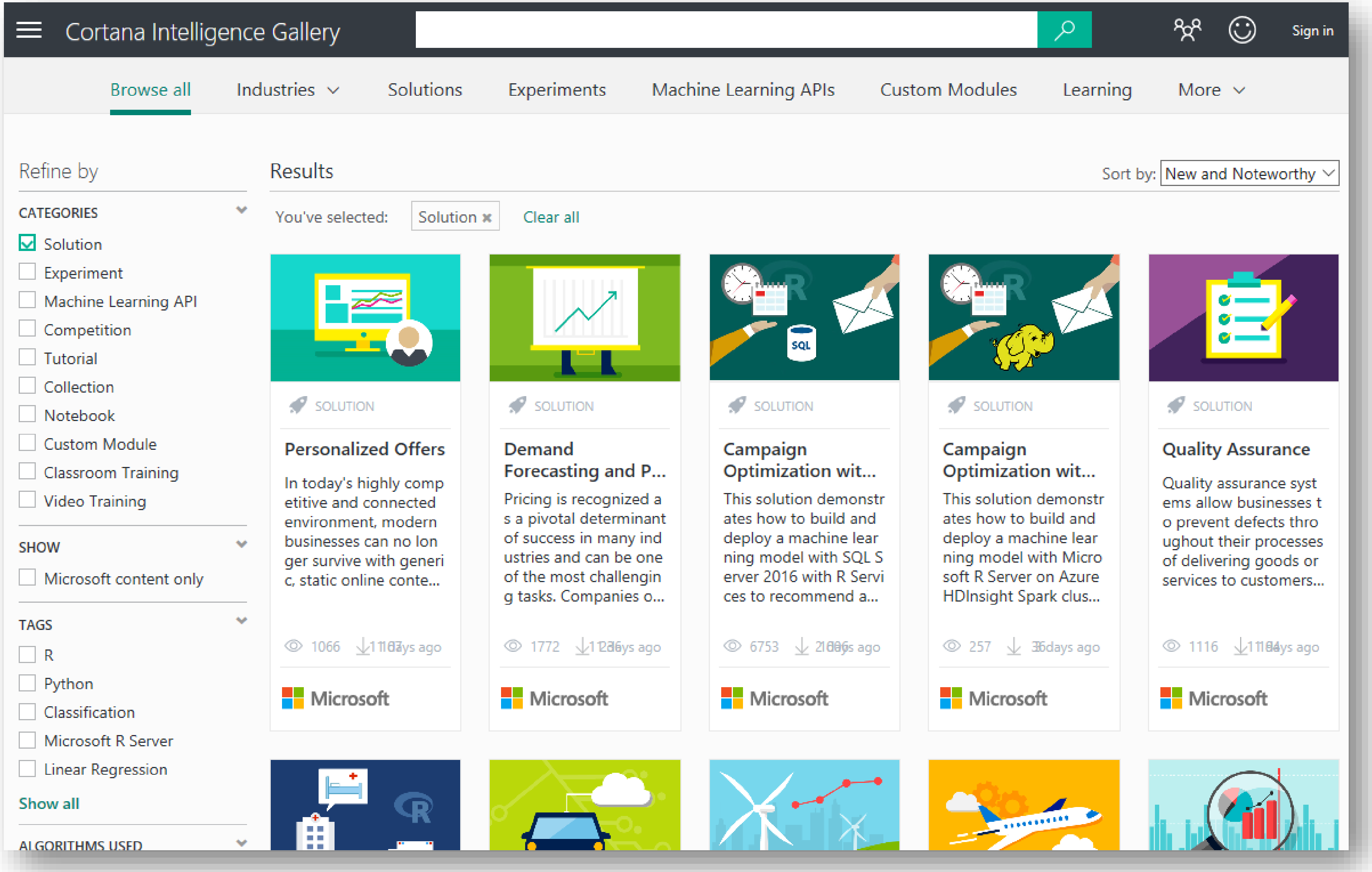
= 139,000 / min

= 2,330 / sec

Web Search Telemetry – Events/Sec – By Hour



Where can I find *even more examples* of this stuff?



☑ Solution
□ Experiment
□ Machine Learning API
□ Competition
□ Tutorial
□ Collection
□ Notebook
□ Custom Module
□ Classroom Training
□ Video Training

□ Microsoft content only

□ R
□ Python
□ Classification
□ Microsoft R Server
□ Linear Regression

Results

Sort by: New and Noteworthy

You've selected: Solution ✖ Clear all

Personalized Offers

In today's highly competitive and connected environment, modern businesses can no longer survive with generic, static online conte...

1066 ↓1116days ago

Demand Forecasting and P...

Pricing is recognized as a pivotal determinant of success in many industries and can be one of the most challenging tasks. Companies o...

1772 ↓1236days ago

Campaign Optimization wit...

This solution demonstrates how to build and deploy a machine learning model with SQL Server 2016 with R Services to recommend a...

6753 ↓2086s ago

Campaign Optimization wit...

This solution demonstrates how to build and deploy a machine learning model with Microsoft R Server on Azure HDInsight Spark clus...

257 ↓36days ago

Quality Assurance

Quality assurance systems allow businesses to prevent defects throughout their processes of delivering goods or services to customers...

1116 ↓1184ys ago

https://gallery.cortanaintelligence.com/browse?categories=["10"]&orderby=freshness desc

Agenda

- What’s next? Wrap up and summary

What’s next for the data platform?

…and what does this mean for us Data Professionals?

1. **On-prem** hosted/deployed data solutions are **diminishing**

2. **Public cloud** data ecosystem is **mature** and **expanding**

3. **IaaS is popular, PaaS is King** (ie *Serverless world is the future*)

4. **Customer “expectation”…**

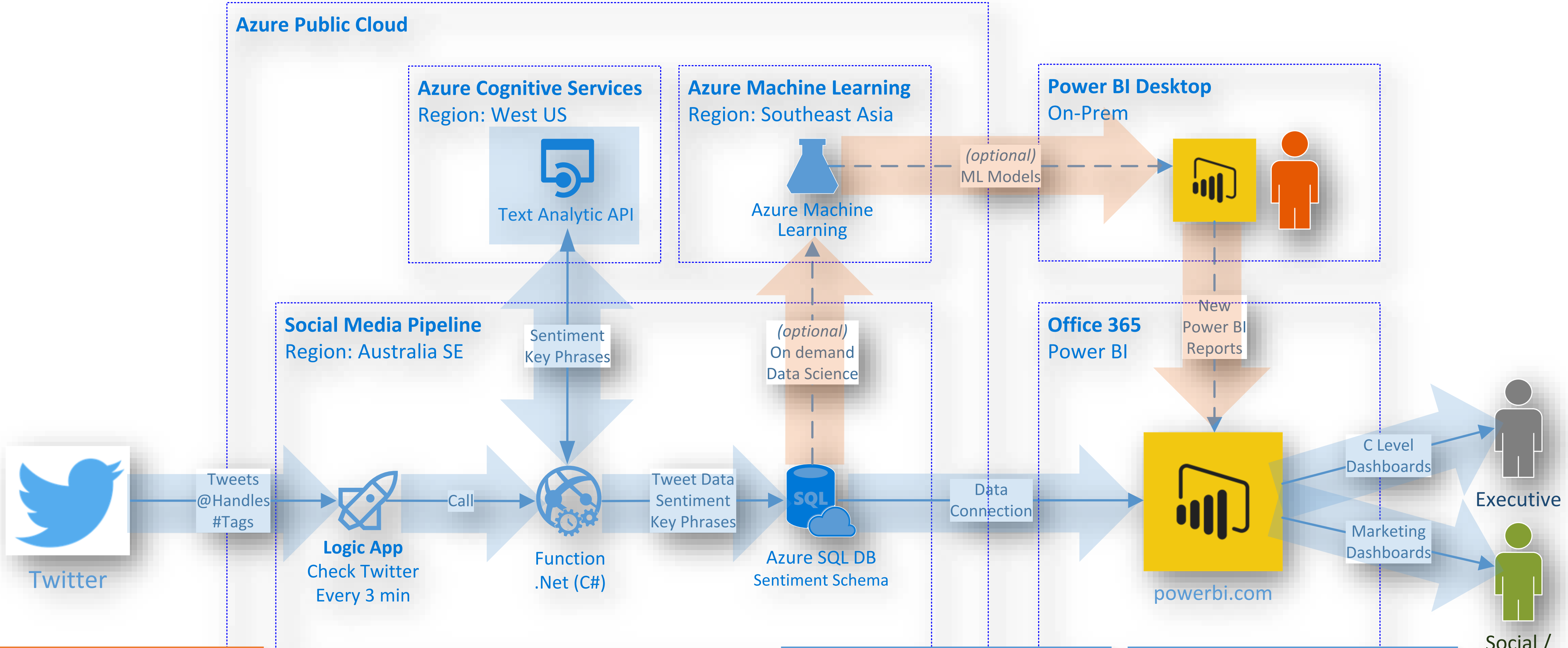
   …This is the “Domain of the Data Professional”

Where can I try this out – or learn more?

- **Cortana Intelligence Gallery Pre-Built Solutions (one-click deploy)**
  - **Vehicle Telemetry**
    https://gallery.cortanaintelligence.com/Solution/Telemetry-Analytics
  - **Personalised Offers**
    https://gallery.cortanaintelligence.com/Solution/Personalized-Offers-2
  - **Energy Demand Forecasting**
    https://gallery.cortanaintelligence.com/Solution/Demand-Forecasting-3

- **EdX Self-Paced Courses (3-4 hrs/week for ~4 weeks)**
  - **Developing IoT Solutions with Azure IoT**
    https://www.edx.org/course/developing-iot-solutions-azure-iot-microsoft-dev225x
  - **Processing Real-Time Data Streams in Azure**
    https://www.edx.org/course/processing-real-time-data-streams-azure-microsoft-dat223-2x-0
  - **Orchestrating Big Data with Azure Data Factory**
    https://www.edx.org/course/orchestrating-big-data-azure-data-microsoft-dat223-3x-0

Your Homework → Twitter Social Media Analytics



This is your homework assignment for today!

Deploys via wizard
Takes ~10 mins

~USD $6 / day
100K tweets / month

https://powerbi.microsoft.com/en-us/solution-templates/brand-management-twitter/

Appendix

APPENDIX AND REFERENCES

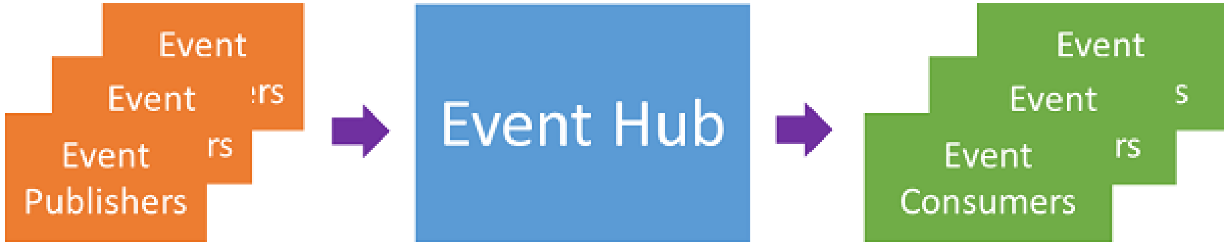
- Online Azure Interactive Services Diagram - http://azureplatform.azurewebsites.net/en-us/
- Azure Time Series Insights Service - https://azure.microsoft.com/en-au/blog/announcing-azure-time-series-insights/
- Service Bus Explorer – Github - https://code.msdn.microsoft.com/windowsapps/Service-Bus-Explorer-f2abca5a
- Cortana Intelligence Gallery - https://gallery.cortanaintelligence.com/
- Predictive Maintenance - https://docs.microsoft.com/en-us/azure/machine-learning/cortana-analytics-playbook-predictive-maintenance
- Anomaly Detection - https://docs.microsoft.com/en-us/azure/machine-learning/machine-learning-apps-anomaly-detection-api
- EdX - Developing IoT Solutions with Azure - https://www.edx.org/course/developing-iot-solutions-azure-iot-microsoft-dev225x
- EdX - Processing Real-Time Data in Azure - https://www.edx.org/course/processing-real-time-data-streams-azure-microsoft-dat223-2x-0
- EdX - Orchestrating Big Data with ADF - https://www.edx.org/course/orchestrating-big-data-azure-data-microsoft-dat223-3x-0
- Microsoft Lambda Architecture - https://social.technet.microsoft.com/wiki/contents/articles/33626.lambda-architecture-implementation-using-microsoft-azure.aspx
- Microsoft Lambda Reference Architecture - https://azure.microsoft.com/en-au/updates/microsoft-azure-iot-reference-architecture-available/
- Wiki Lambda Architecture - https://en.wikipedia.org/wiki/Lambda_architecture
- Azure Stream Analytics Query Language - https://msdn.microsoft.com/en-us/library/azure/dn834998.aspx
- Azure Stream Analytics Query Windowing Functions - https://msdn.microsoft.com/en-us/library/azure/dn835019.aspx
- Azure Stream Analytics Query Patterns - https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-stream-analytics-query-patterns
- Azure Storage Explorer - http://storageexplorer.com/

Azure Event Hubs - *on one slide*

INFORMATION MANAGEMENT
https://azure.microsoft.com/en-us/services/event-hubs/

- **Fully Managed Service (PaaS) for ingesting events/messages at a massive scale (*think telemetry processing from websites, IoT etc*).**

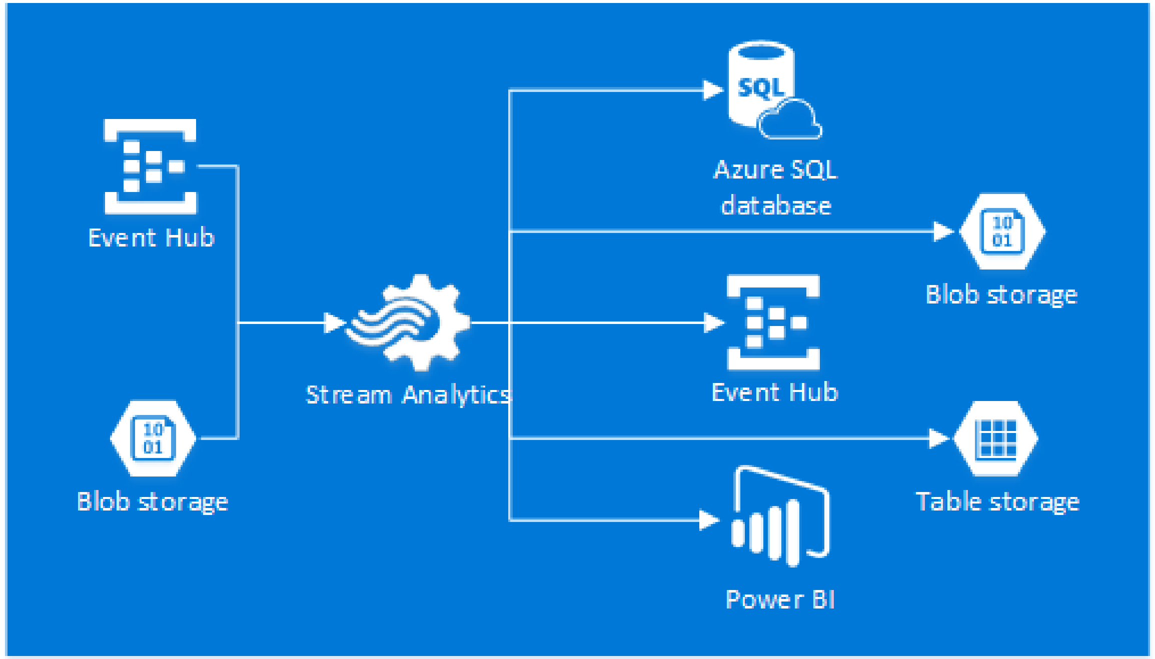
- **Acts as the “front door” to high velocity data traffic**
  - An event ingestor sits between event publishers and consumers
  - Allows asynchronous decoupled solutions to be architected



Azure Stream Analytics - *on one slide*

ANALYTICS
https://azure.microsoft.com/en-us/services/stream-analytics/

- **Fully Managed Service (PaaS) for deploying CEP solution/services**
- CEP = Complex Event Processing = high scale event ingestion and in-stream analytics
- Transform, augment, correlate, temporal operations, reference data
- SQL-like Language to perform in-stream queries and produce tabular result sets



Azure Data Lake Store & Analytics - *on one slide*

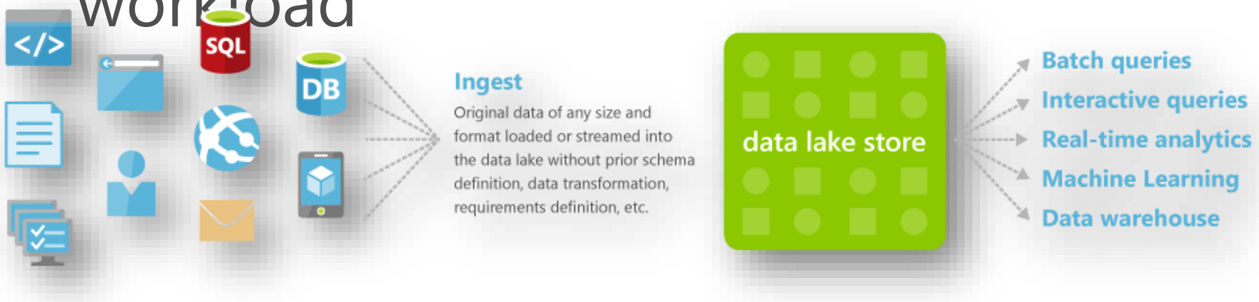
BIG DATA STORES

https://azure.microsoft.com/en-us/solutions/data-lake/

https://azure.microsoft.com/en-us/services/data-lake-analytics/

Azure Data Lake - Store

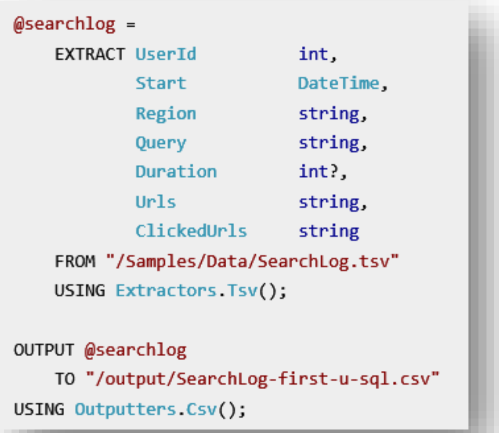
- PaaS service, nothing to manage
- Highly scalable distributed file store
- Unlimited storage, PB size files
- Capture data of any size or shape
- Tuned for analytic/streaming workload



SQL
DB
Ingest
Original data of any size and format loaded or streamed into the data lake without prior schema definition, data transformation, requirements definition, etc.
data lake store
Batch queries
Interactive queries
Real-time analytics
Machine Learning
Data warehouse

Azure Data Lake - Analytics

- PaaS service, nothing to manage
- Introduces new language called U-SQL
- Build batch jobs to process data
- Dynamic scaling of job performance
- Integrates with Azure services



```
@searchlog =
    EXTRACT UserId          int,
            Start           DateTime,
            Region          string,
            Query           string,
            Duration        int?,
            Urls            string,
            ClickedUrls     string
    FROM "/Samples/Data/SearchLog.tsv"
    USING Extractors.Tsv();

OUTPUT @searchlog
    TO "/output/SearchLog-first-u-sql.csv"
USING Outputters.Csv();
```

Azure SQL Data Warehouse - *on one slide*

BIG DATA STORES

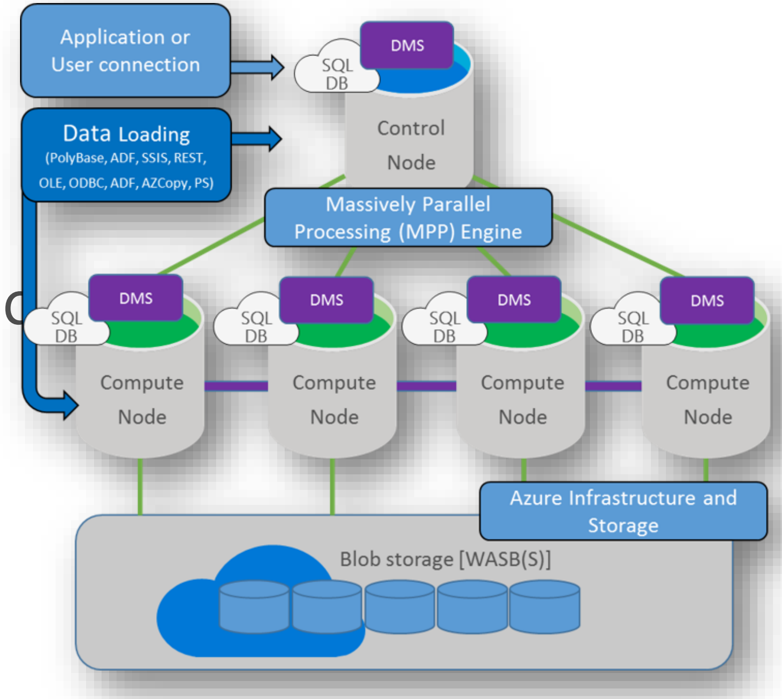
https://azure.microsoft.com/en-us/services/sql-data-warehouse/

https://en.wikipedia.org/wiki/Massively_parallel_(computing)

- **Fully Managed Service (PaaS) for deploying an MPP SQL Data Warehouse**
- Is essentially deploys distributed Azure SQL Databases under the hood
- Is an Azure cloud version of on-prem SQL Server APS

- **Compute**
  - Leverages MPP technology to provide scale
  - Dynamically scale compute resource up to 20x on demand
  - Pause compute resource on demand to reduce costs

- **Storage**
  - Massive storage scale up to a PB SQL relational data


Application or User connection
Data Loading
(PolyBase, ADF, SSIS, REST, OLE, ODBC, ADF, AZCopy, PS)
SQL DB
DMS
Control Node
Massively Parallel Processing (MPP) Engine
SQL DB
DMS
Compute Node
SQL DB
DMS
Compute Node
SQL DB
DMS
Compute Node
SQL DB
DMS
Compute Node
Azure Infrastructure and Storage
Blob storage [WASB(S)]

Azure Machine Learning - *on one slide*

ANALYTICS

https://azure.microsoft.com/en-us/services/machine-learning/

- **Fully Managed Service (PaaS) for composing analytical models for performing predictive analytics.**
- Drag/drop GUI interface to create and deploy predictive solutions
- Can integrate with Azure to source data, and write outputs
- Lots of pre-configured solutions can be deployed from the ML gallery

Classification *

- Assign a **category** to each item (i.e. tweet data sentiment analysis)

Regression *

- **Predict** a **real value** for each item based on **features** (i.e. predict house sale price) ☺

Clustering *

- **Partition** items into homogeneous **groups** (i.e. finding similar companies based on characteristics)

Azure Cognitive Services API’s

Give your solutions a human side

https://www.microsoft.com/cognitive-services/en-us/documentation

Microsoft Cognitive Services preview

Vision (5)

Computer Vision | Emotion | Face | Video | Moderator

Speech (3)

Custom Recognition | Speaker Recognition Speech

Language (6)

Bing Spell Check | Translator | Language Understanding Linguistic Analysis | Text Analytics | Web Language Model

Knowledge (5)

Academic Knowledge | Entity Linking | Q&A Maker | Knowledge Exploration | Recommendations

Search (5)

Bing Auto Suggest | Bing Image Search | Bing News Search Bing Video Search | Bing Web Search

Azure Data Factory - *on one slide*

INFORMATION MANAGEMENT

https://azure.microsoft.com/en-us/services/data-factory/

- **Fully Managed Service (PaaS) for Composing Data Processing, and Movement Services into Scalable and Reliable Data Pipelines.**

- **Access Data Sources (source and target)**
  - Many supported data sources – not as many as SSIS but growing
  - https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-movement-activities/
- **Perform Data Transformation (in the pipeline)**
  - Hive, Pig, MapReduce, Azure ML, SQL Stored Proc, ADL U-SQL, .Net (…*and growing!*)
  - https://azure.microsoft.com/en-us/documentation/articles/data-factory-data-transformation-activities/

Azure Data Catalog (ADC)

What is it?

Fully managed cloud metadata repository service

Discover, catalog and make searchable various business data sources

Manage the process of locating and securely consuming those sources

Crowdsource annotation of the data source, tables/objects and columns

Simple to use web interface for registering and managing data sources

ADC keeps track of the data sources, it DOES NOT hold the data!

What can you do with it (Use Cases)

Want to centrally register all relevant business data sources

Self-Service BI and providing power users a central point to locate the data they need

Capturing tribal business data knowledge (crowdsourcing data documentation)

Azure CosmosDB (DocDB) (NoSQL) (PaaS)

NoSQL document database-as-a-service (**PaaS**), managed by Microsoft Azure.

Native support for **JavaScript**, **SQL** and txns over schema-free **JSON** documents

[JSON = JavaScript Object Notation]

Built for cloud-designed apps

- Write **procedures**, **triggers** and **UDF's** using **JavaScript**
- **Reliable** and **predictable** performance, **scale up** on demand
- Automatic **geo-redundant** data copies, automated **backup**

Rich Query and Transactions over Schema-free Data

Query schema-free data (agile development)

Native JavaScript transactional processing

Familiar SQL-based query language

Reliable & Predictable Performance

Fast, predictable performance

Tunable consistency

Elastic scale (massive scalability)

Rapid Development

Build with familiar tools – REST, JSON, JavaScript

Easy to start and fully-managed

Enterprise-grade Azure platform