Microsoft **Azure Data Catalog** (ADC)

Data Management and Integration in the **Cortana Intelligence Suite**
https://azure.microsoft.com/en-us/services/data-catalog/

Rolf Tesmer
Data Solutions Architect
Microsoft

About me…

- **Data Solution Architect** – (Anything in Azure that is **data** related)
- About **20 years** IT experience
- Interest in **Azure Data Services**, **data warehousing**, **visualization**, **spatial**
- **Am on LinkedIn *here*:** https://www.linkedin.com/in/**rolftesmer**
- **Am I blog *here*:** https://**mrfoxsql**.wordpress.com/

MY assumptions about **YOU**!

- Have an interest in **data**
- Awareness of **Azure Cloud**
- Limited exposure to **Cortana**

Agenda

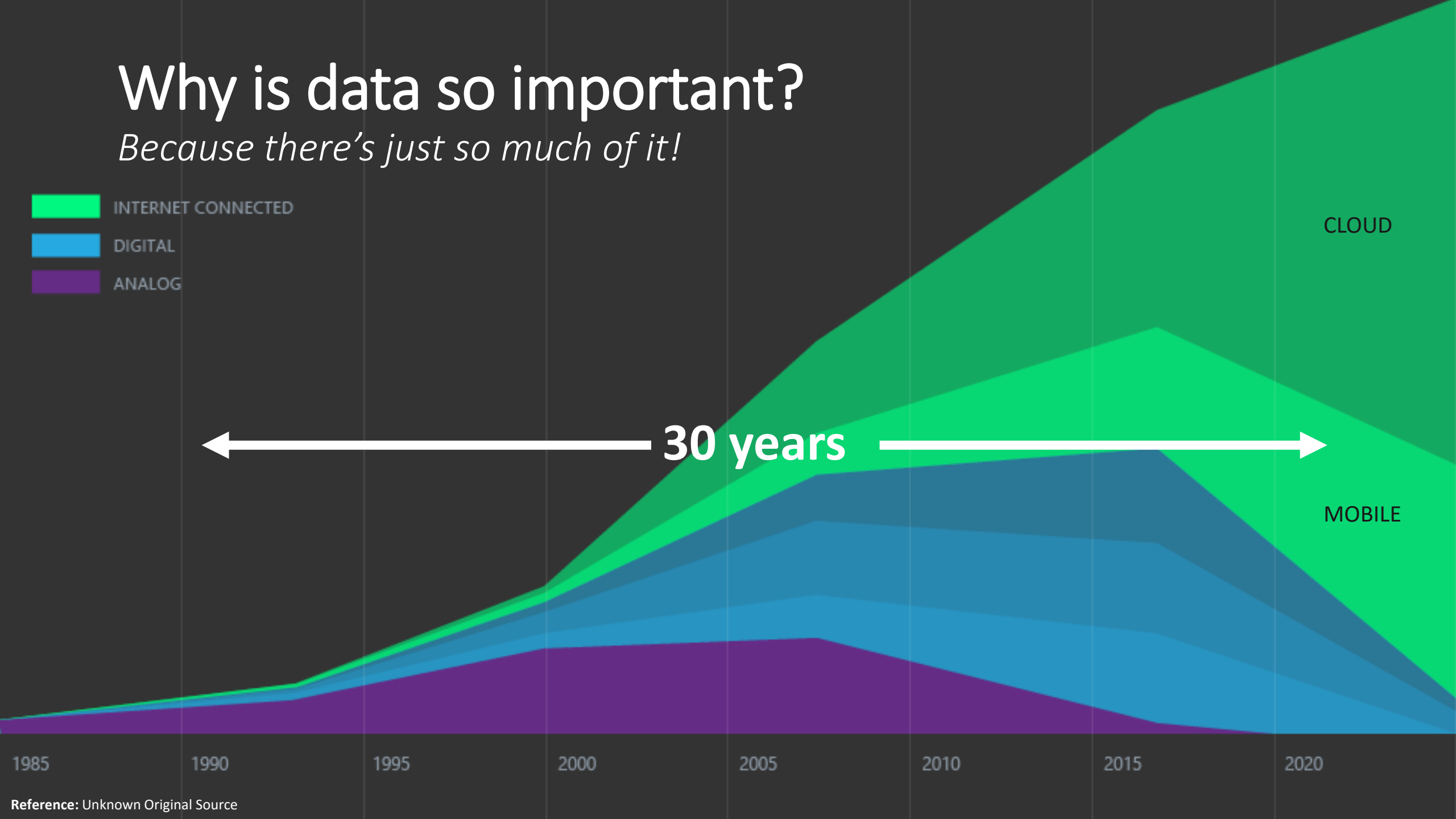
1	Introduction
2	Key Components of the Microsoft Azure Cortana Intelligence Suite
3	Azure Data Catalog – End to End
4	Q & A

Agenda

1 Introduction

Why is data so important?

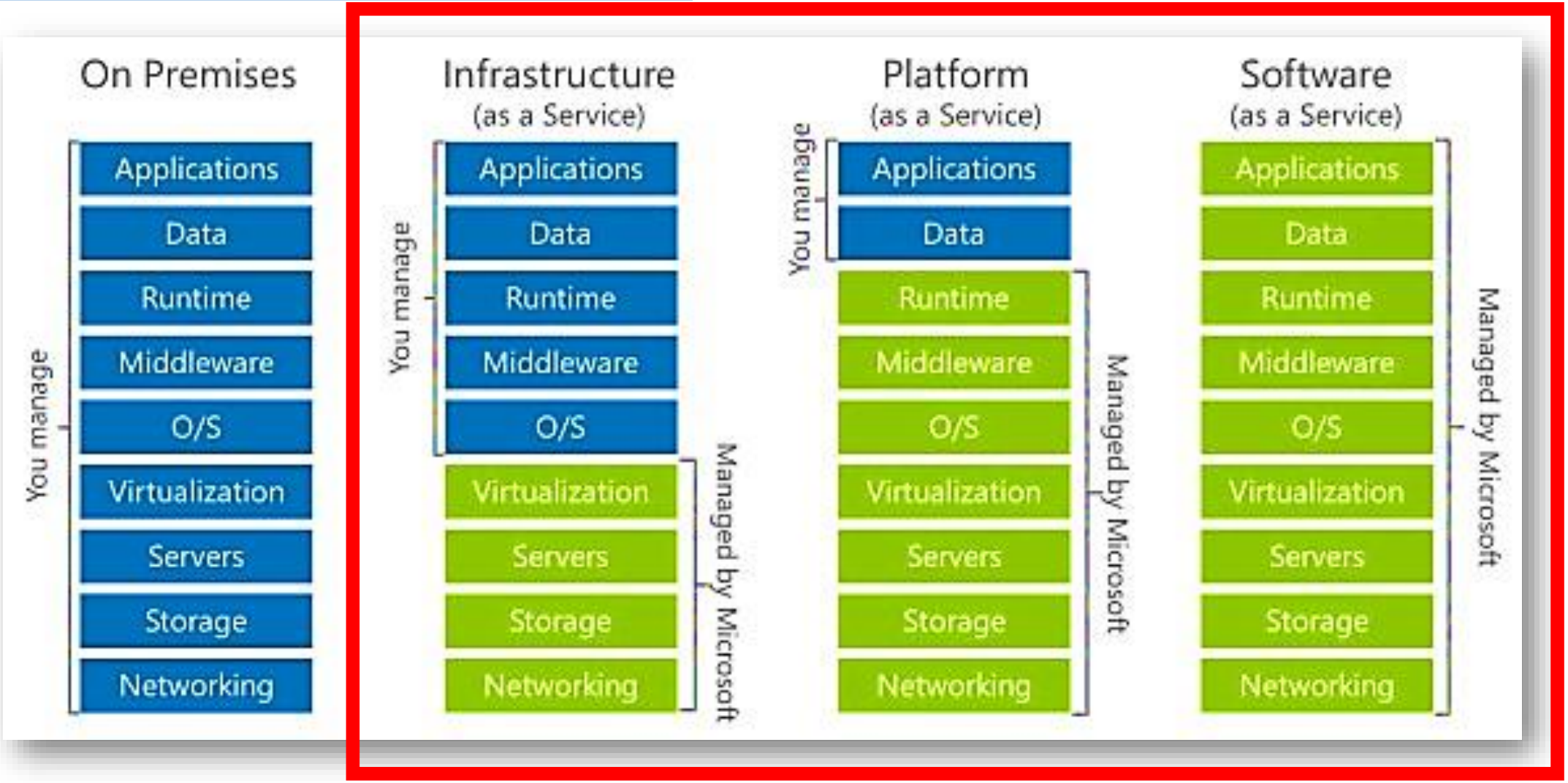
Because there's just so much of it!



Reference: Unknown Original Source

Platforms; On-Prem vs Cloud

https://azure.microsoft.com/en-us/overview/what-is-iaas/
https://azure.microsoft.com/en-us/overview/what-is-paas/
https://azure.microsoft.com/en-us/overview/what-is-saas/



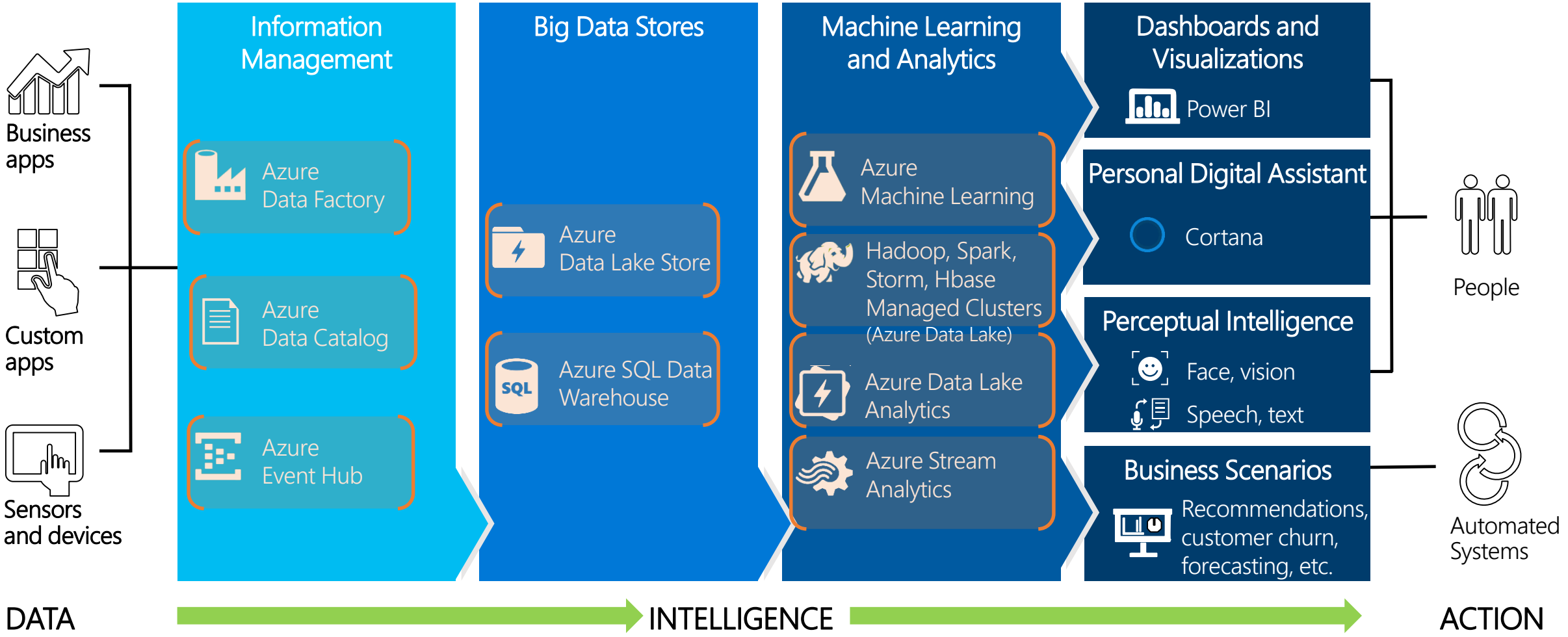
Reference: Unknown Original Source

Agenda

2 Key Components of the Microsoft Azure Cortana Intelligence Suite

Cortana Intelligence Suite

Transform data into intelligent action



Cortana Intelligence Services Diagram - https://www.microsoft.com/en-us/cloud-platform/what-is-cortana-intelligence-suite
All Azure Services Interactive Diagram - http://azureplatform.azurewebsites.net/en-us/

An Example Cortana Intelligence Deployment…

Streaming

Batch

Devices and Services (Streaming)

Phones

Application Services

Messages / Telemetry

Internet Services

Data Stores and Databases (Batch)

Databases

FIles

Azure Cortana Deployments

Catalog

Azure Data Catalog

Event Hubs

Stream Analytics

Data Lake

Data Lake Analytics

HDInsight

Data Factory

SQL Data Warehouse

Reporting & Analytics

Machine Learning

Power BI

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 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 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

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();
```

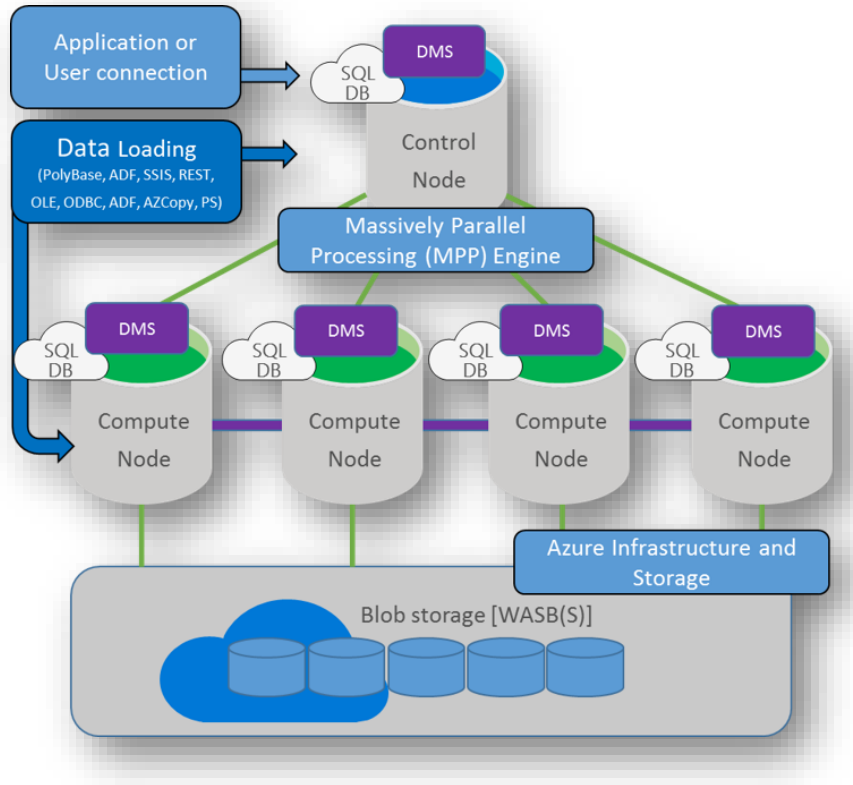
Azure SQL Data Warehouse - *on one slide*

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

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 HD Insight - *on one slide*

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

- **Fully Managed Service (PaaS) for deploying Hadoop, Spark, HBase and Storm**
- Available on Windows and Linux
- 100% OPEN SOURCE Apache Hadoop (HDP 2.3) compatible

HIVE
- **HiveQL** is a **SQL-like** language (subset of SQL) (Compiled into **MapReduce** jobs)

HBASE
- **Columnar**, **NoSQL** database on data in **HDFS**

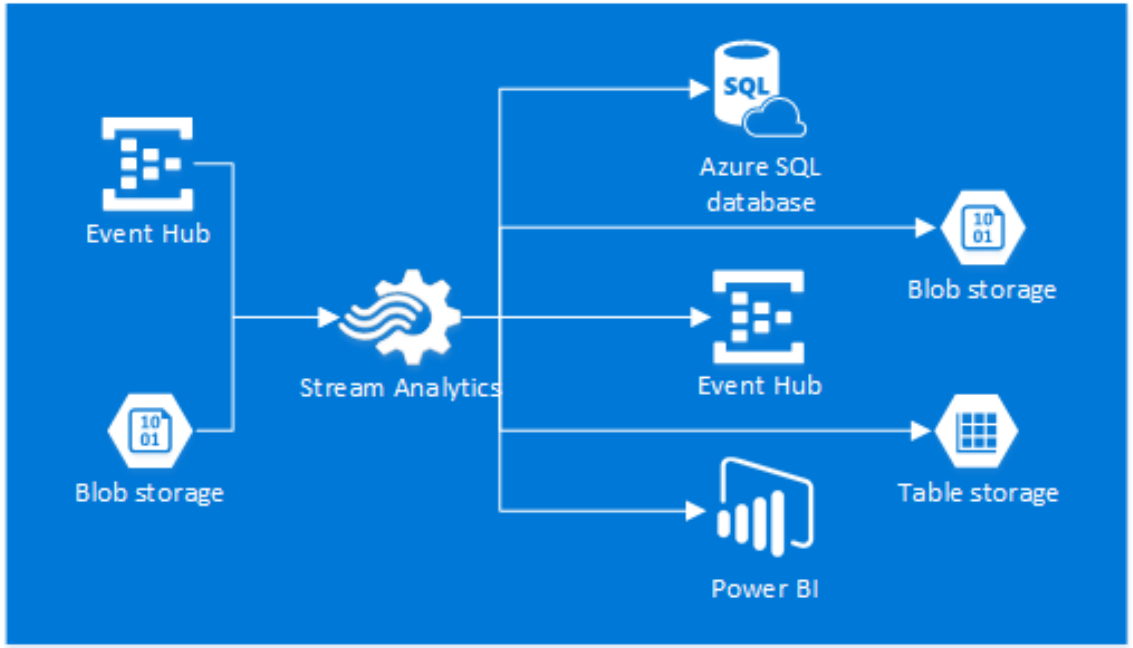
SPARK
- **In Memory** Processing on Multiple Workloads

STORM
- **Stream Analytics** for Near-Real Time Processing (similar to Azure Stream Analytics)

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



Agenda

3 Azure Data Catalog – End to End

Overview – Do you have any of these Challenges?

I spend more time **looking** for data, than I do actually analyzing it!

Our data is sitting in **multiple sources**, but I don't know which data sits where!

We have many different **data ecosystems** across the enterprise, but we have no way to share data artifacts across them!

Our users are busy **re-producing** data assets that they don't know already exist!

We have no way of **tracking** our BI and Analytics assets

Overview - What is Azure Data Catalog (ADC)?

A metadata repository that allow any user to register, enrich, understand, discover, and consume data sources

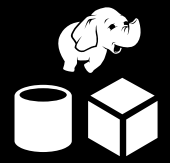
An enterprise-wide catalog in Azure that enables self-service discovery of data from *any source*

Overview - How is ADC Different?

Data ***source*** discovery

Metadata only – Data resides where it lives. No need for data movement, No latency

Data from ***any*** source



Structured and unstructured on premises and in the cloud

Consumption through ***any*** tool

Enabling publishing, discovery and consumption of data sources through your tool of choice

Powered by annotation ***crowdsourcing***

Empowering any user to capture and share their knowledge about registered sources

Overview – Who Uses ADC?

Publisher
(Data Steward)

Publish
Register Data Sources
Manage Metadata

Enrich
Categorize – Annotate

Consumer

Discover
Browse – Search – Consume

Enrich
Categorize – Annotate

Understand
Get context – Identify Intent

IT Admin

Govern
Apply Policies - Control Access

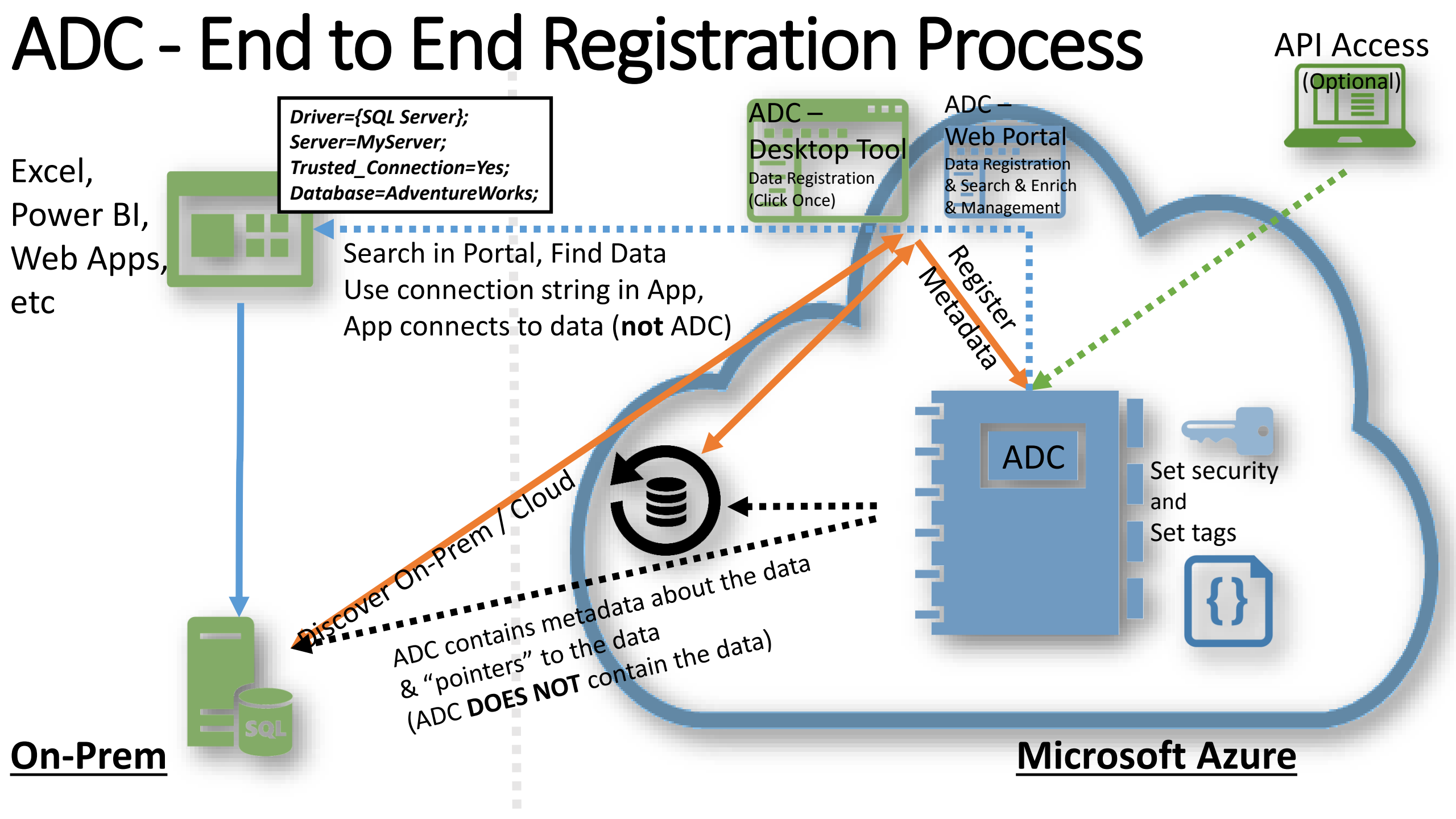
Analyze
Track and monitor usage

ADC - Key Terms

https://azure.microsoft.com/en-us/documentation/articles/data-catalog-terminology/

- **Catalog** cloud based metadata repository holding **data assets**
- **Data Source** a system containing the **data asset** to be catalogued
- **Data Asset** objects in a data source that can be catalogued
- **Registration** the process of extracting and storing **data asset** information
- **Structural Metadata** data describing the physical structure of a **data asset**
- **Descriptive Metadata** data describing the purpose or intent of a **data asset**
- **Data Preview** snapshot of 20 records from certain **data asset** types
- **Data Profile** snapshot of the **data asset** table/column structural definition
- **Expert** a user of that **data asset** who is recognised as an expert
- **Owner** a user who has full rights over a **data asset**

ADC - End to End Registration Process



Demo

1. Browse to portal home page - **explain what we see**
2. View and interact with data assets - **explain what we see**
3. Create and share new data asset – **Desktop Tool**
4. Create and share new data asset – **Manual Entry**
5. Search for data assets - **explain what we see**
6. Consume data sets – **Excel top 1,000 rows**

Implementing an Enterprise ADC

https://azure.microsoft.com/en-us/documentation/articles/data-catalog-adopting-data-catalog/
https://azure.microsoft.com/en-us/documentation/articles/data-catalog-get-started/
https://feedback.azure.com/forums/34192--general-feedback

- **Need to approach like any project – but be aware of data specific focus**

  - **Prepare a Vision Statement –**
    - What is the business problem? What are the goals? What are the timelines?

  - **Identify Stakeholders and Roles –**
    - **People** - Executive Sponsor, Influencers/Champions, Pilot Group, End Users
    - **Roles** - Owners, Experts, Data Stewards, Users
    - Understand if there will be cultural impacts or blockers to adoption

  - **Define a Pilot Project**
    - **Team** – Target the right team that has a specific, targeted/focused and measurable need
    - **Facilitation** – define and communicate a plan/timeline, create a project buzz and excitement
    - **Training** – Ensure there is a structured approach to initial awareness and education
    - **Tracking** – continually revisit the progress against executive/sponsor expectations

  - **[Play-Repeat] for all aspects of the business**

Agenda

4 Q & A

…Q & A

Thank You…