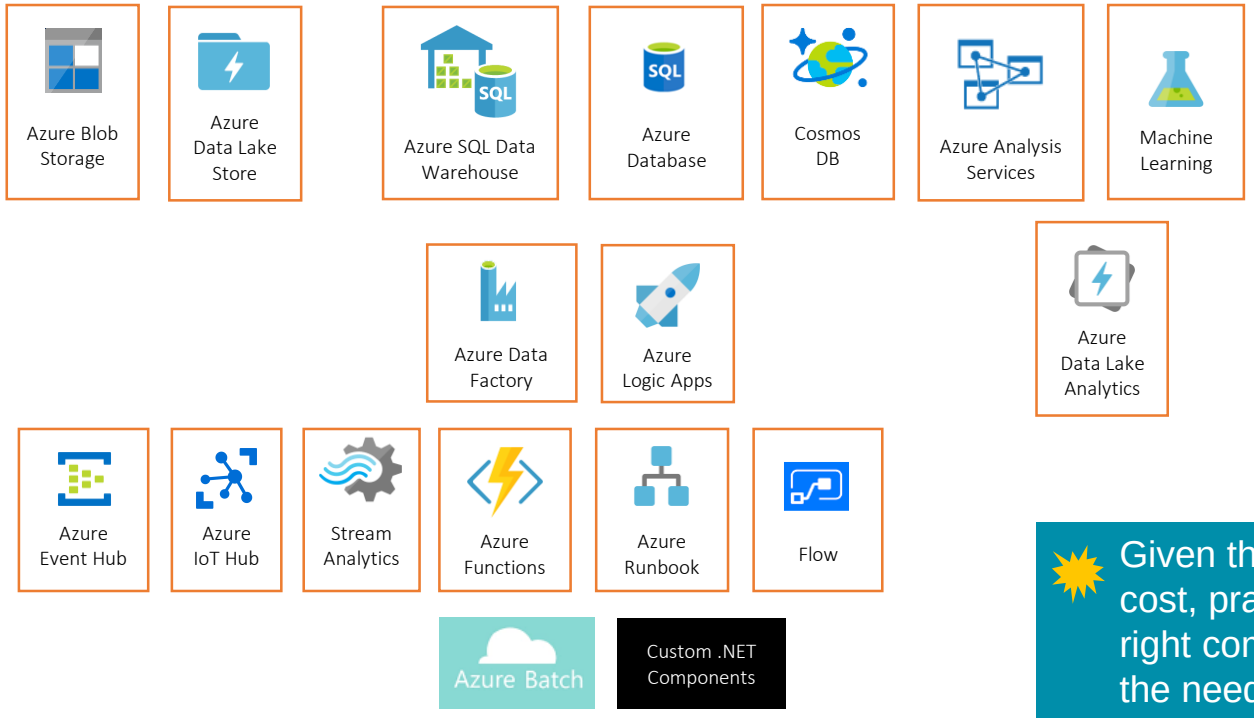
How to Architect Azure for Insights & Analytics

Mehmet Bakkaloglu
Business Insights & Analytics, Hitachi Consulting
SQLBits Conference, London, Feb 2018

Vast array of Platform-as-a-Service components

HITACHI
Inspire the Next



Given the differences in functionality, cost, practicality, how to choose the right components, bearing in mind the needs of the business now and in the future?

HITACHI
Inspire the Next

This is a wide topic that depends on the business and type of data…

So we will use a real-world Azure BI implementation over 2 years as a vehicle for this discussion, and look at the decisions made along the way…

Contents

HITACHI
Inspire the Next

- Evolution of Azure Architecture at a Water Company client:

Phase 1
- Basic Architecture
- Orchestration tools
- Storage options

Phase 2
- IoT data: Near real-time analytics
- Web App for data entry

Phase 3
- Annotations & Reference Data Management
- Notifications
- Connect to non-on-premise sources

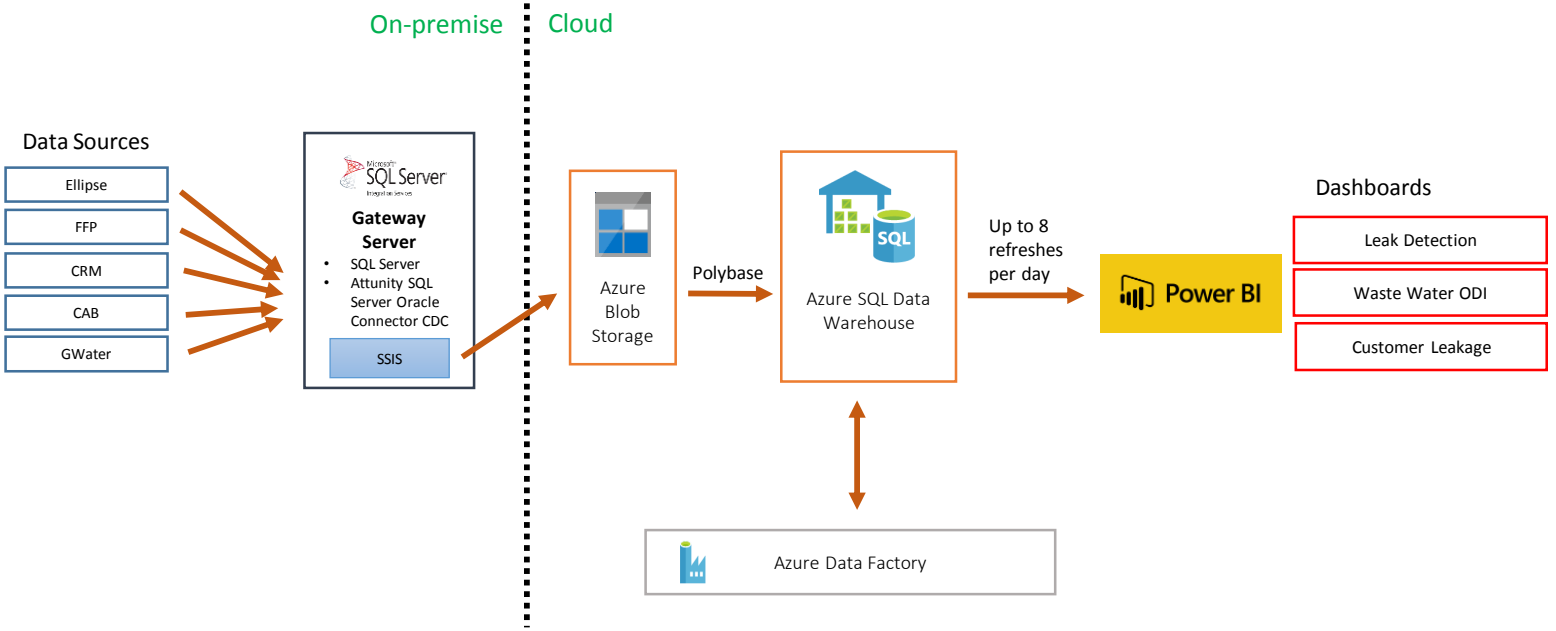
Further decisions
- Environments
- Coordination (multiple ELT on Azure DW)
- PowerBI options

Phase 1 (Feb 2016 – Jul 2016)

First Phase: Basic Architecture

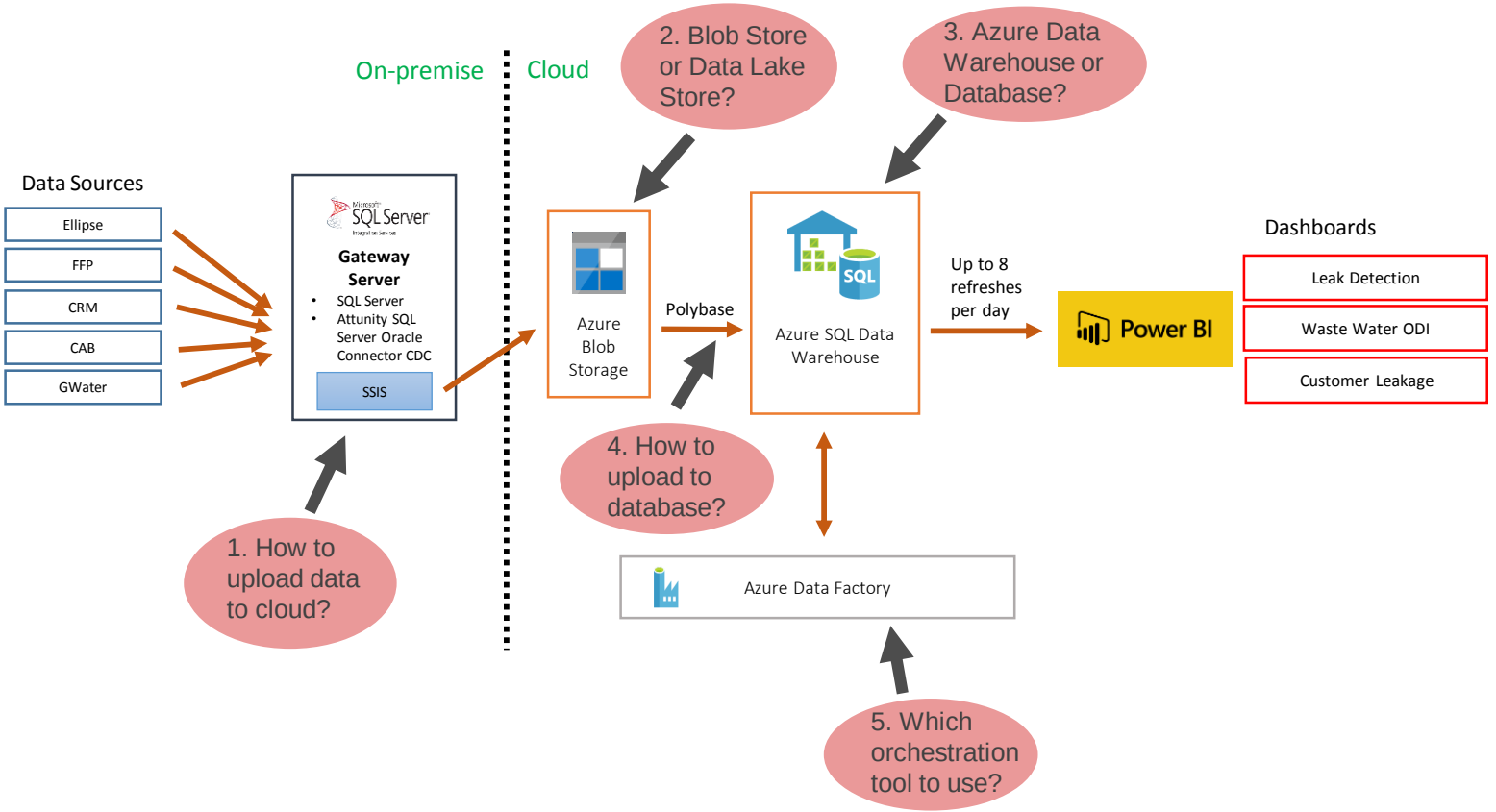
Feb 2016 – Jul 2016

HITACHI
Inspire the Next



Key Decisions

HITACHI
Inspire the Next



1. How to upload data to Azure

HITACHI
Inspire the Next

- SSIS – Azure Blob Upload Task / Azure Data Lake Store File System Task

  \+ Can connect to various sources, do transformations, conditional tasks etc.

  – Burden of having to manage software / server

- ADF Copy

  \+ PaaS

  – Data Management Gateway required – may not be allowed by some organisations due to security restrictions

  – It's a tool for data movement and orchestration – some limitations in ability to implement logic (ADF v2 has more features)

- Other options:
  - AZCopy – command-line utility
  - BCP – suitable for small files
  - Custom .NET App
  - Other proprietary software (e.g. Mulesoft)

2. Blob Storage vs Data Lake Store

HITACHI
Inspire the Next

- Blob Storage – cheap general purpose storage
- Data Lake Store
  - Optimised for large analytics workloads – log files, IoT data, social media, clickstream
  - Built-in Hadoop support
  - Can be used with U-SQL + C#
  - Authentication based on Azure Active Directory
  - Processed data may be dumped into Azure Data Warehouse

  (Note: Data Lake became generally available in Nov 2016, hence we did not use it in that particular instance)

	Blob Storage	Data Lake Store
1 TB & 1,000 read/write	£17/month	£70/month
10 TB & 10,000 read/write	£170/month	£700/month
100 TB & 100,000 read/write	£1700/month	£7000/month

Note: Exact price depends on usage. Prices are North Europe region on 11/02/2018.

Data Lake Store price quadruple!

If you have a mixture of general purpose files and large analytics data, it may be better to go with Data Lake

3. Azure Data Warehouse vs Azure Database

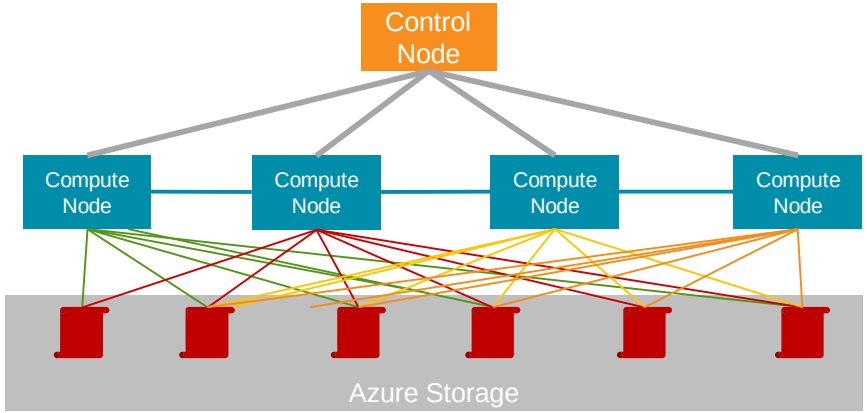
HITACHI
Inspire the Next

Azure Data Warehouse	Azure Database
MPP System, for processing very large amounts of data	Typical SQL Server, SMP System
Can be scaled up/down, can be paused	It can be scaled up as well (though less flexible) but can't be paused
High price And with Dev/Test/Prod environments price easily goes up Min 100 DWU, 1TB: £857/month 400 DWU, 10 TB: £4,035/month 6000 DWU, 240TB: £69,593/month	Low price Basic 5 DTU, 2GB: £3.6/month S2 50 DTU, 250GB: £55/month S12 3000 DTU, 1TB: £1,700/month
Max 240TB (Permanent database tables)	Max 4TB (some regions 1TB)
32 concurrent queries, 1024 concurrent sessions	Max concurrent queries (6400) and concurrent sessions (30000)
Should not be accessed by web apps -- Only for OLAP	Can be accessed by web apps / websites -- can be used for OLTP or OLAP
Polybase for loading data fast	No Polybase, but there is Bulk Insert

Prices are North Europe region on 11/02/2018

Azure Data Warehouse

HITACHI
Inspire the Next



- MPP System
- Control Node distributes work to the Compute Nodes
- Data Movement Service: Movement of data required between Compute Nodes

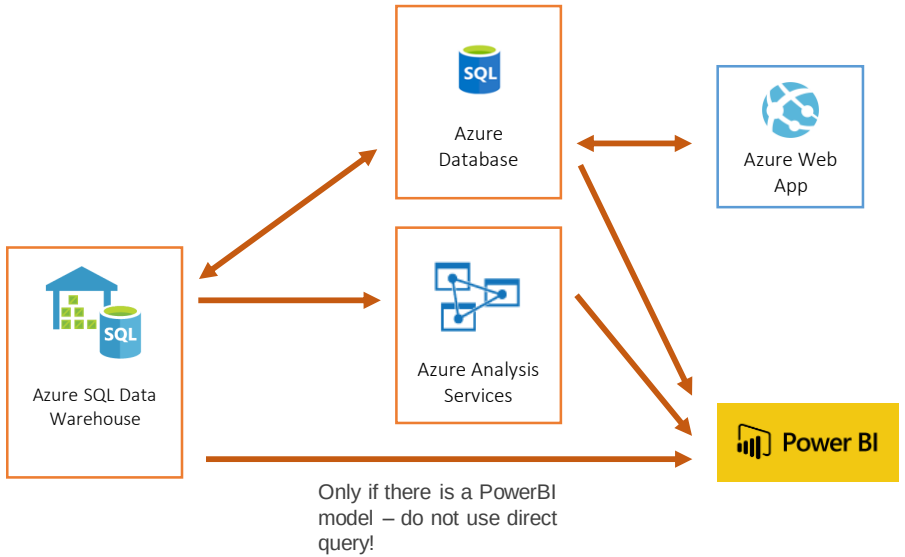
Azure Data Warehouse vs Azure Database

HITACHI
Inspire the Next

Only go for Azure Data Warehouse if there is a large amount of data to be processed

Note: Ideally need to decide upfront whether Azure Data Warehouse is needed, as can't easily migrate between the two, lots of differences in functionality – CTAS, Primary Keys / Foreign Keys, PolyBase etc.

If we go with Azure Data Warehouse, it is quite likely Azure Database will be required for certain operations later:



4. How to upload data to Azure DW

HITACHI
Inspire the Next

- PolyBase is the fastest method!
  - With PolyBase throughput increases when DWU is increased
  - This is not the case for other methods such as BCP, ADF, SSIS, Custom C#

```
CREATE EXTERNAL TABLE [stg].[Product]
([ProductKey] [int] NOT NULL,
[ProductName] nvarchar(100) NULL )
WITH (LOCATION='/Product/',
DATA_SOURCE = AzureBlob,
FILE_FORMAT = TextFileFormat,
REJECT_TYPE = VALUE (Or PERCENTAGE),
REJECT_VALUE = 10);
```

- But:
  1. Beware of number of files when using PolyBase
     - If there is a huge file, performance slow, so split into small files
     - If 100,000s files (e.g. IoT devices producing files at short intervals with only a few lines), performance suffers. So:
       a) Merge small files before loading
       b) Archive files
     - If number of files reasonable, PolyBase performance is super!
  2. Can't do row-level error logging with PolyBase – can only specify what number / percentage of rows to ignore
     - So you may either use a different method or split out the error rows before loading
  3. Can only load UTF-8 & UTF-16 files

4. How to upload data to Azure Database

HITACHI
Inspire the Next

There is no Polybase – so:

From Blob:

- Azure Data Factory
- Custom C#
- BULK INSERT T-SQL

```
BULK INSERT TableName
FROM 'data/filename.dat'
WITH ( DATA_SOURCE = 'AzureBlobStorageAccountName');
```

- OPENROWSET table-value function

```
SELECT Field1, Field2
FROM OPENROWSET(BULK 'data/ filename.dat', DATA_SOURCE =
'AzureBlobStorageAccountName',
        FORMATFILE='data/filename.fmt',
        FORMATFILE_DATA_SOURCE = '
AzureBlobStorageAccountName') as TableName;
```

From Data Lake:

- Azure Data Factory
- Custom C#

5. Which orchestration tool to use?

HITACHI
Inspire the Next

Azure Data Factory

- Out-of-the-box connectors, parallelisation, logging, ability to run stored procedures etc
- Can use custom .NET components if needed
- In Visual Studio need to code in JSON or can use UI through Azure Portal

SSIS with Azure Data Factory v2

- Out-of-the-box connectors in SSIS
- UI
- Can use custom.NET components
- Useful if SSIS packages need to be migrated to cloud (packages automatically upgraded to latest version when deployed to Azure)

.NET with Web Job

- Flexibility
- But need to do everything yourself, connectors, parallelisation etc.
- Dependent on .NET Framework

SSIS on Azure VM

- IaaS (need to maintain software)
- Out-of-the-box connectors in SSIS
- UI
- Can use custom components
- Useful if SSIS packages need to be migrated to cloud

Phase 2 (Aug 2016 – Feb 2017)

Question #1

HITACHI
Inspire the Next

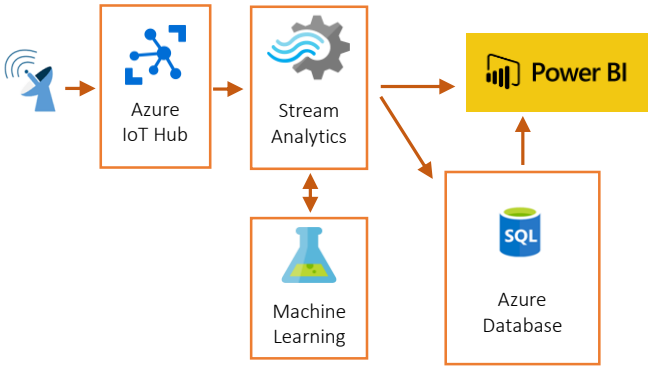
How to do near real-time analytics on IoT data in Azure?

a) Stream Analytics
b) Apache Storm on Azure HDInsight
c) Another option

Stream Analytics

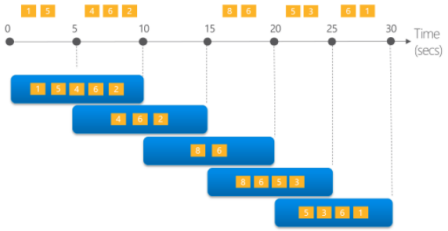
HITACHI
Inspire the Next

Typical architecture:



Window functions:

- Hopping
- Tumbling
- Sliding



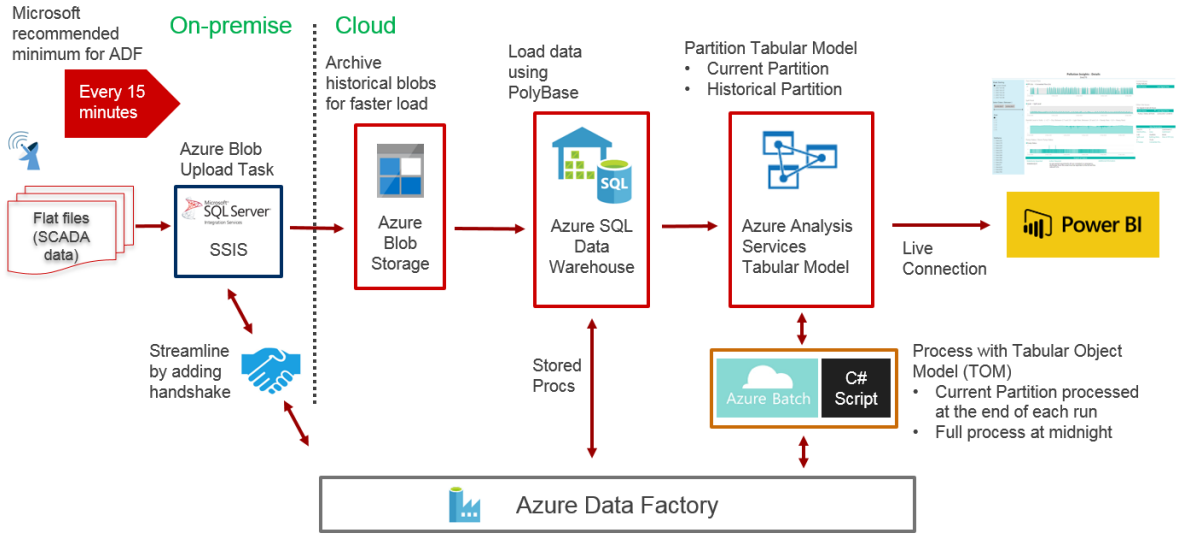
https://docs.microsoft.com/en-us/azure/stream-analytics/stream-analytics-window-functions

- Issues to consider:
  - It provides window functions and can connect to Azure Machine Learning but **for complex analytics data has to be dumped into database / data warehouse**
  - **In real world often we cannot connect to IoT devices** – because service is provided by third parties or devices are hidden behind a demilitarized zone (e.g. national infrastructure) – **so all we get is regular batch files**

➢ Therefore stream analytics may not always be the best option…

Alternative solution for near real-time analytics

HITACHI
Inspire the Next



Process only a stream of data every 15 min

At midnight process monthly / historical partition

- 3 key points to achieve near real-time:
  - Partitioned Tabular Model – process only current day partition during the day
  - Archive historical blobs (if files are small and many, may also need to merge files before uploading to Blob)
  - Handshake between SSIS and ADF
- Added benefit of doing historical analytics alongside near real-time

  (More details on this in my SQLBits 2017 presentation)

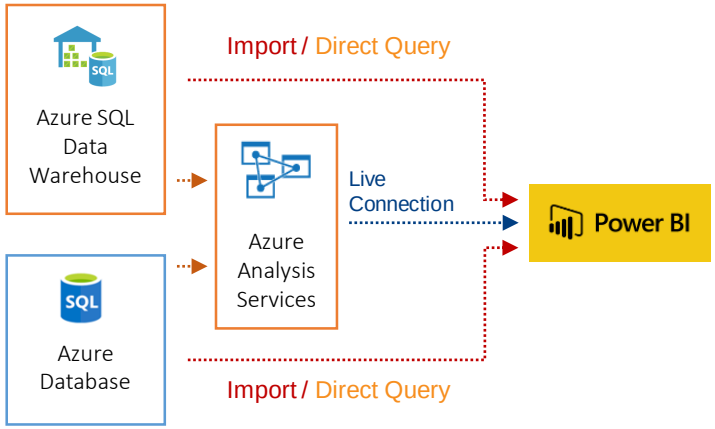
Side Question

HITACHI
Inspire the Next

If we were starting fresh, when would we use Azure Tabular Model, when we would use just PowerBI?

PowerBI Model vs Azure Tabular Model

HITACHI
Inspire the Next



PowerBI Connection Options

Import Mode: Data is imported into PowerBI – can be refreshed up to 8 times per day – has data volume and performance restrictions

Direct Query: No data is imported into PowerBI – refresh not required – has performance restrictions

Live Connection: Connect to Analysis Services where modelling and calculations are done – provides near real-time reporting – best for large amounts of data and high performance

	PowerBI Model	Azure Analysis Services Tabular Model
Pros	• Low Cost • With Import mode modelling functionality similar to Analysis Services	• No size limit • Can implement near real-time refreshes by partitioning • Can scale up / down
Cons	• Direct Query is slow so can't be used in practice • In Import mode can only be refreshed 8 times per day with Pro license *(With PowerBI Premium this is unlimited, but there is additional cost to that)* • File size 10GB with Pro license *(With PowerBI Premium it is 10GB too)* • Cannot be partitioned	• Cost is high – use it only when it's really required! B1: ~£234/month, S1: ~£1,104/month • Additional development time and management of Azure Analysis Services Prices are North Europe region on 11/02/2018

Question #2

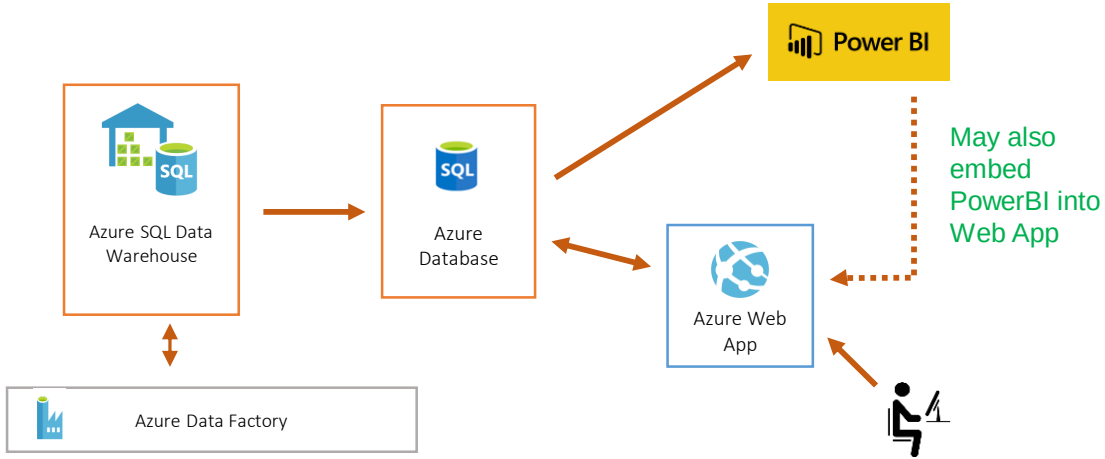
HITACHI
Inspire the Next

How to build a reporting system where users enter / view data using a complex Web App?

User Interaction

HITACHI
Inspire the Next

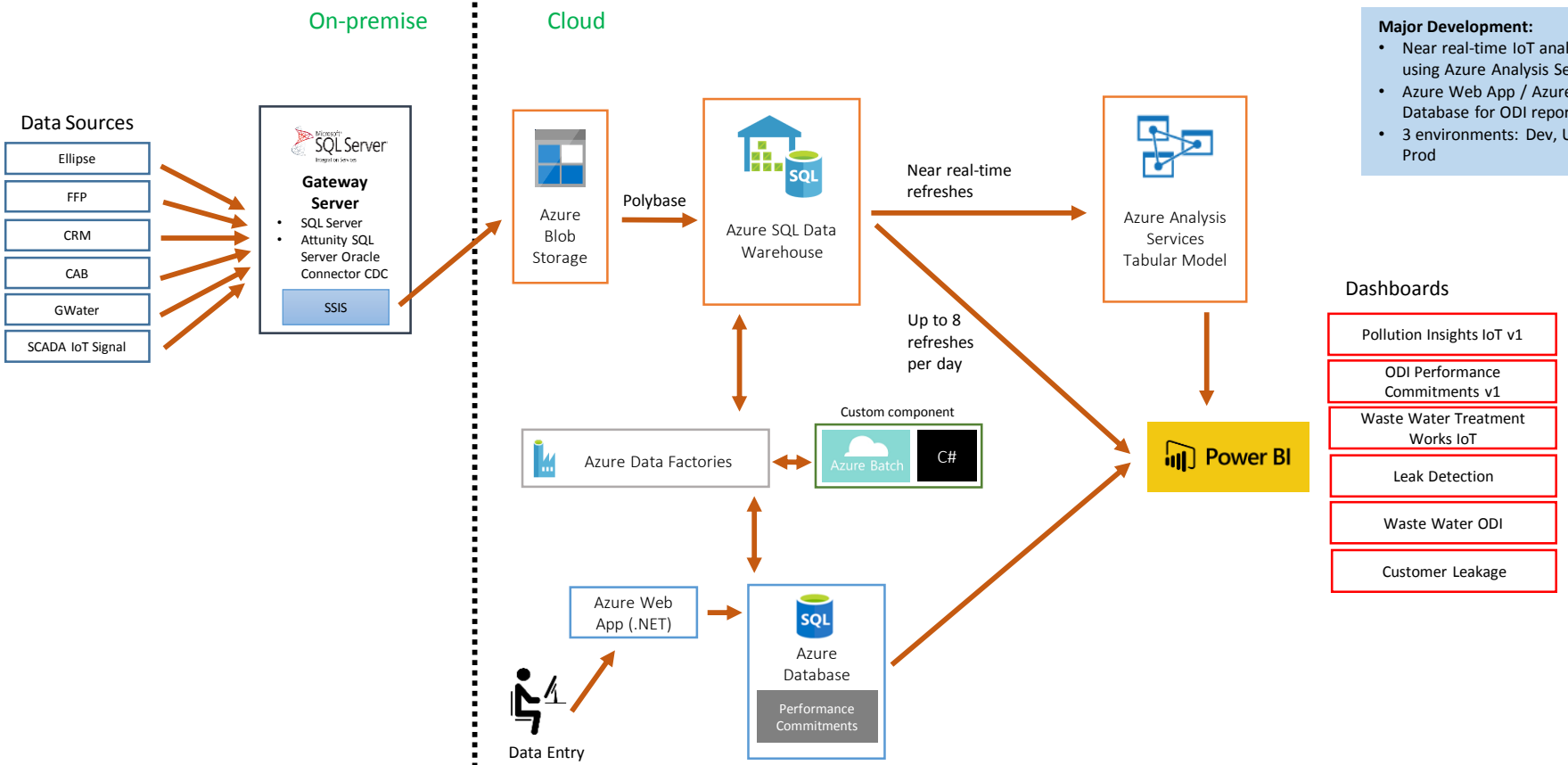
- Web App should only interact with Azure Database
- Summarised data from Azure Data Warehouse may be dumped into Azure Database



Second Phase

Aug 2016 – Feb 2017

HITACHI
Inspire the Next



Phase 3 (Mar 2017 – Dec 2017)

Questions

HITACHI
Inspire the Next

1. How do we do annotations & reference data management?
2. How do we do notifications?
3. So far we had only on-premise data sources… How do we upload data from sources such as web services and SFTP?

1. Annotations / Reference Data Management

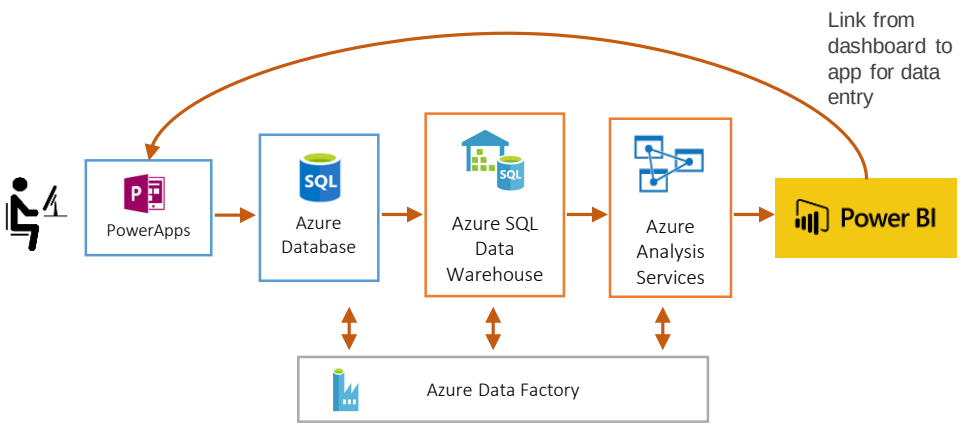
HITACHI
Inspire the Next

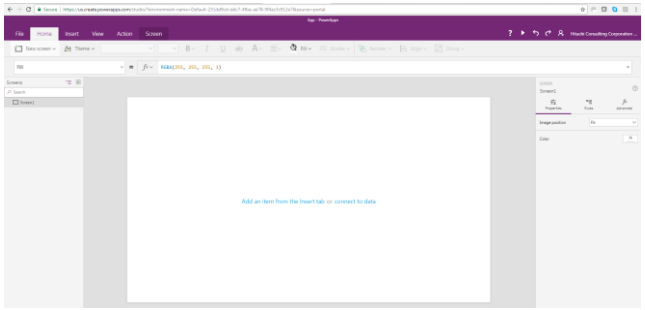
a) PowerApps

- Use it if a simple form is needed for annotations (insert, update, delete records from a table) or reference data management
- Quick development time through web interface
- As of January **PowerApps Custom Visual for Power BI** is available so form can be embedded in dashboard (However, it is still not possible to trigger refresh of dashboard through form)

b) Custom Web App

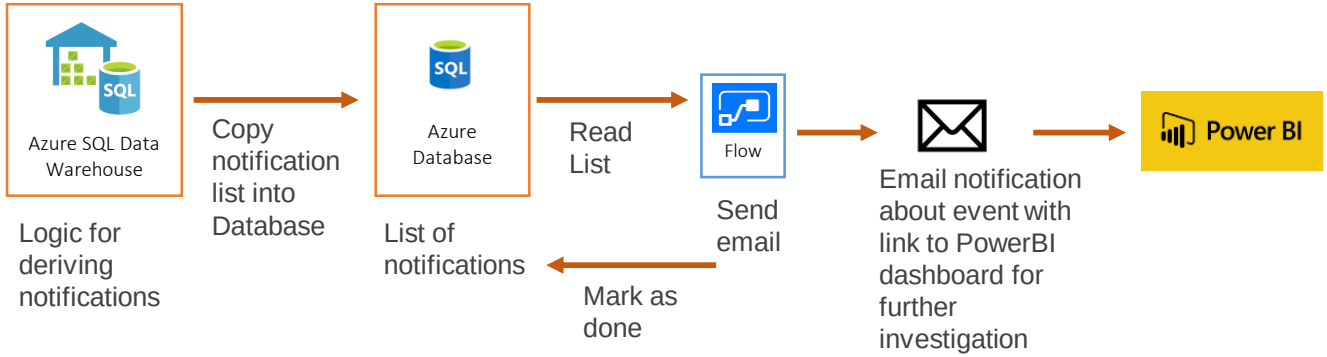
- For more sophisticated apps



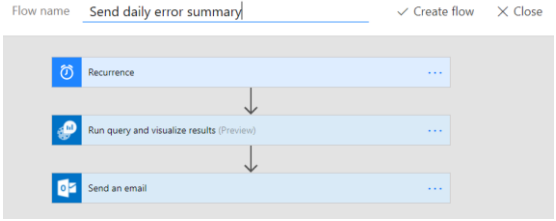


2. Notifications

HITACHI
Inspire the Next



- With **Microsoft Flow** it is easy to do
- Alternatively write custom .NET App



3. Uploading data from non-on-premise sources

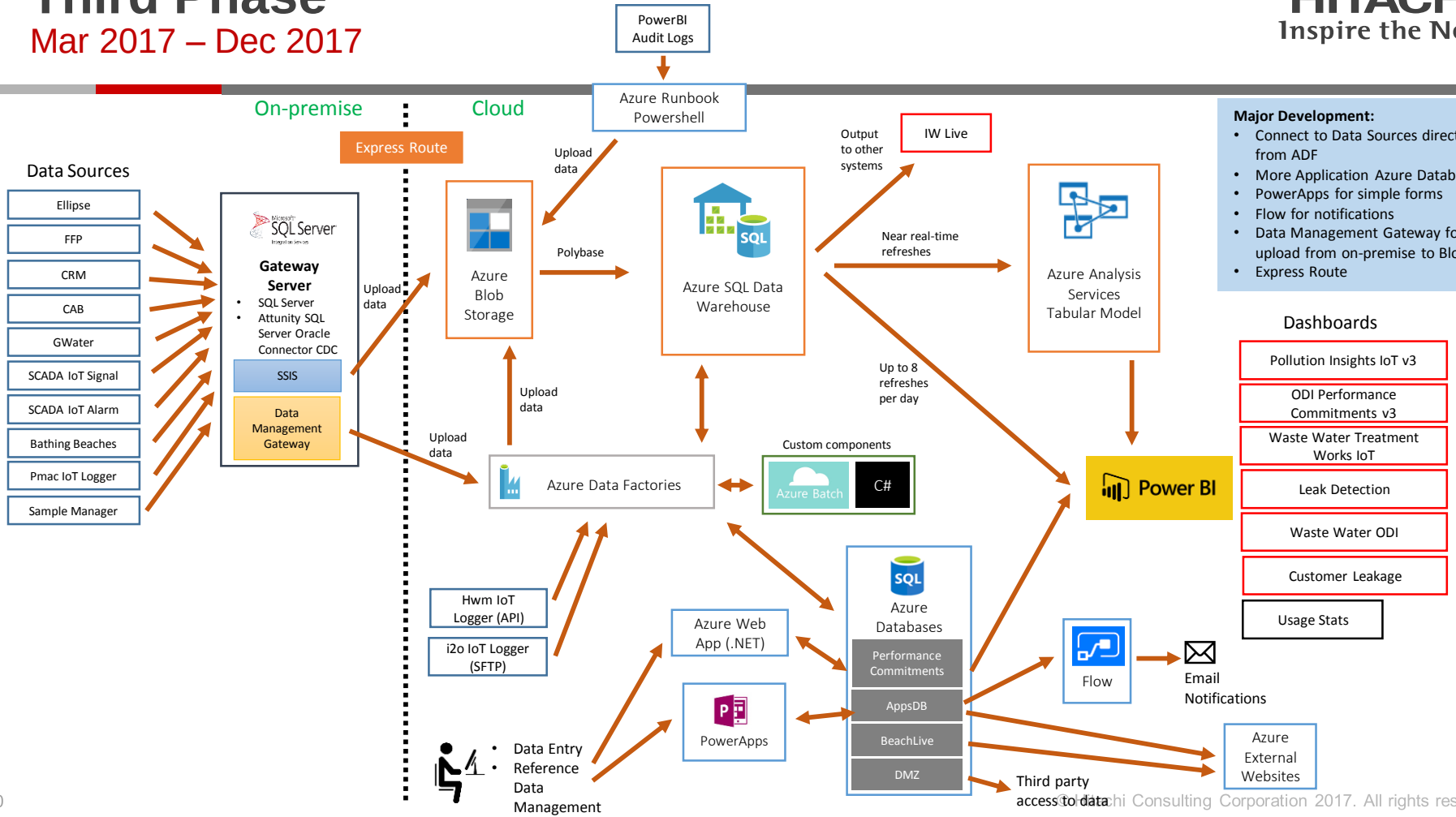
HITACHI
Inspire the Next

- Not much point in landing the data on premise
- Can use Azure Data Factory
  - Connectors: Web services, SFTP, Amazon Redshift, Oracle, SAP etc.
- Otherwise use custom .Net code
  - Run using ADF + Azure Batch, or Web Job
- Or a third party tool (e.g. Mulesoft)

Third Phase

Mar 2017 – Dec 2017

HITACHI
Inspire the Next



Major Development:

- Connect to Data Sources directly from ADF
- More Application Azure Databases
- PowerApps for simple forms
- Flow for notifications
- Data Management Gateway for data upload from on-premise to Blob
- Express Route

Further Decisions

Further Decisions

HITACHI
Inspire the Next

1. How do we coordinate multiple Azure Data Factories, Tabular Models, PowerBI dashboards accessing Azure DW *when it is scaling up and down, and given that only 32 multiple concurrent connections are allowed*?

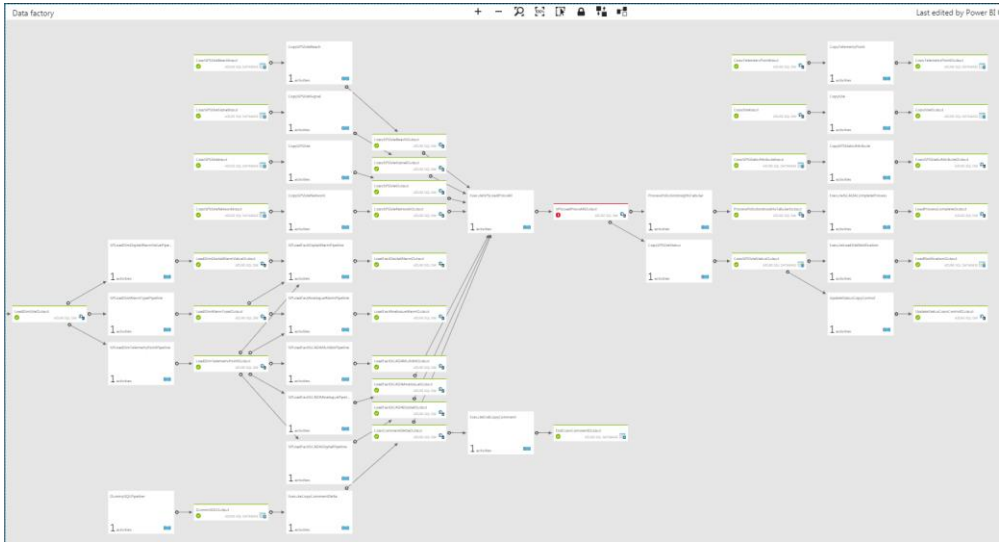
2. How do we manage Dev-Test-UAT-Prod environments?

3. When do we need PowerBI Premium?

Azure Data Factory

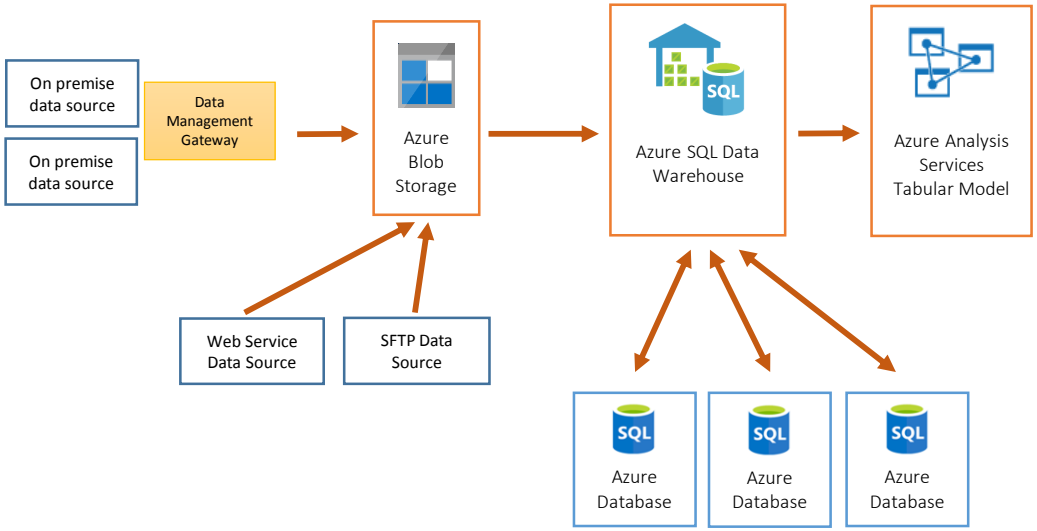
HITACHI
Inspire the Next

- ADF used for data movement and orchestration
- Example ADF loading facts and dimensions, and processing tabular model (it is possible to simplify this by having a master proc to load facts and dimensions but then will not be parallelised):



Typical Azure Data Factory Operations

HITACHI
Inspire the Next



- As solution grows, operations required may comprise of:
  - Process data in Azure Data Warehouse
  - Load data from on-premise to Blob Storage
  - Load data to Blob from sources such as web services, sftp, Azure Databases
  - Copy processed data from Azure Data Warehouse to Azure Database to be accessed by Web App
  - Process Azure Analysis Services Tabular Model
  - Scale up/down Azure Data Warehouse
  - Etc

→ So we will end up with multiple ADFs…

1.a. How to coordinate multiple ADFs?

HITACHI
Inspire the Next

In the basic architecture single ADF will be sufficient

As solution grows, single ADF difficult to manage, therefore will need multiple ADFs

How to coordinate multiple ADFs?

- No functionality to have a master ADF

  → Therefore functionally ADFs need to be distinct

- But if one ADF scales Azure DW up / down, Azure DW becomes unavailable, resulting in failures in other ADFs accessing Azure DW. So:

  a) Either implement long retries on Azure DW operations so there is an interval between retries

```
"policy":
{
    "retry": 3,
    "longRetry": 5,
    "longRetryInterval": "00:10:00"
}
```

  b) Or use Azure Automation Runbook to stop/restart ADFs

1.b. Azure Data Warehouse – Resource Classes

HITACHI
Inspire the Next

- Maximum concurrent queries in Azure DW is 32!
  - As number of concurrent operations on Azure Data Warehouse increases (i.e. multiple Data Factories, Tabular Model refreshes), to ensure predictable performance will need to manage resource allocation
- Assign admins/users to Resource Classes
  a) Dynamic Resource Classes (smallrc, mediumrc, largerc, xlargerc): allocate memory depending on current DWU
  b) Static Resource Classes: allocated fixed memory

Best plan depends on what complete solution looks like
→ e.g. ADF solutions may run under a admin account with a large resource class, and Tabular Model refreshes may be done with a different admin account with a large resource class, to ensure they don't block each other

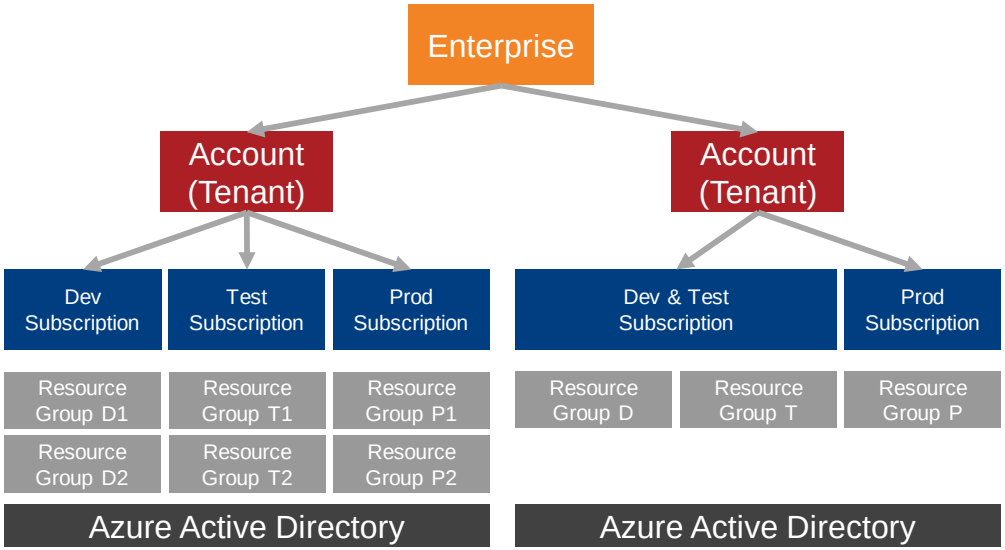
DWU	Maximum concurrent queries	Concurrency slots allocated	Slots used by smallrc	Slots used by mediumrc	Slots used by largerc	Slots used by xlargerc
DW100	4	4	1	1	2	4
DW200	8	8	1	2	4	8
DW300	12	12	1	2	4	8
DW400	16	16	1	4	8	16
DW500	20	20	1	4	8	16
DW600	24	24	1	4	8	16
DW1000	32	40	1	8	16	32
DW1200	32	48	1	8	16	32
DW1500	32	60	1	8	16	32
DW2000	32	80	1	16	32	64
DW3000	32	120	1	16	32	64
DW6000	32	240	1	32	64	128

https://docs.microsoft.com/en-us/azure/sql-data-warehouse/sql-data-warehouse-develop-concurrency

2. Managing Environments

HITACHI
Inspire the Next

Hierarchy:



- Typically an organisation has one tenant

- Within a tenant create subscriptions for each environment (or consider merging Dev with Test)
- Subscriptions allow easy billing and access control

- Resource Group is associated with a subscription
  - Resources in a group should have the same lifecycle
  - All resources may be placed within a group or may be separated
    - e,g. Database and Web App for a specific solution may be placed in a separate group

3. PowerBI Considerations

HITACHI
Inspire the Next

- PowerBI Premium
  - Provides dedicated capacity
  - No restrictions on dataset refresh rate
  - Users who require read-only access (do not do self-service or collaborate)
  - Publish reports on-premise on PowerBI Report Server
  - Cost is high (e.g. around £4K for an organization with 500 users), therefore not suitable for a small organisation
- PowerBI Pro
  - For users who require self-service and collaboration

https://docs.microsoft.com/en-us/power-bi/service-premium-faq

Essentially there are 2 options:

Small-to-medium organisations
Get X number of PowerBI Pro licences, and other users will use free service

Medium-to-large organisations
Get X number of PowerBI Pro licences, and consider getting PowerBI Premium

Prices are North Europe region on 11/02/2018

Other cool services

HITACHI
Inspire the Next

- Azure Automation Runbook: schedule PowerShell and Python scripts to automate tasks
- Logic Apps: Automate your workflows (Flow is built on this, and is a simpler version)
- Azure Function: To run a simple piece of code
- Cognitive Services: Image / Speech / Language processing algorithms to be used in apps and bots
- Azure Data Lake Analytics: Process petabytes of data using U-SQL, R, Python and .NET
- Cosmos DB: NoSQL (e.g. if you have json files)
- HDInsight: Big Data Analytics using open source tech, Hadoop, Spark, Hive, Kafka, Storm, R

Last few things…

HITACHI
Inspire the Next

- Contact: https://www.linkedin.com/in/mehmet-bakkaloglu-0a328010/

- **Hitachi Consulting** is a platinum sponsor of SQLBits – please stop by our stand for more information

- Would appreciate if you could complete the feedback form

Thank you