**Project:**So I will briefly take you through the latest and the recent project which we have been developing.  
Our project revolves around a solution application known as Albert cube, which is a reporting mechanism for displaying the financial data of the ecommerce business unit of Albert Heijn. Whatever transactions happen on a daily basis, items being sold, revenue generated, all this information is available in Albert cube and it is entirely based in Pallas landscape which is Legacy DW. Business used to slice and dice the measures/facts and the dimension using Oracle BI and the cubus infrastructure is hosted on Oracle Essbase.  
We envisioned how to recreate the Albert cube from the data platform and replicate the same functionality in Azure Native PAAS solutions and for that we have taken help of various Azure services like ADF, ASA, Autoloader and Power BI.   
Dimension: Distribution centre(brands on which the product is based on),PUP, supplier & seller information,Time, Promo, delivery,  
Fact: No of articles sold,revenue,discount,VAT, cost price, Margin%,gross profit, no of order

So I will try to briefly explain the overall architecture of the project.  
We are maintaining 3 different zones for our project

**1)Pallas Staging**

**2)Landing Zone**

**3)Hammer solution zone**

**1)Pallas Staging:** on premise legacy DW where the files are being kept in Inbound.

Power jobs(Linux based Unix scripts) will pick up the files from inbound and place them in the mount point of Outbound.

MFT jobs(Managed File Transfer) access the files in the mount point of Outbound(Pallas staging SFTP) and place them in Landing Zone storage account(Data Lake). [MFT is a tool in Webmethods Active Transfer built for configuring file transfer, once data gets placed into Data Lake, it will delete the files from the source SFTP, scheduled every 20 mins, if no file it will wait, no staging, direct load from source to target]

**2)Landing Zone:** LZ is explicitly created for processing the Hammer files, we are getting the fixed with files, so we are using Autoloader to check the availability of files in the LZ/source and immediately writing it into Staging zone, then we are doing basic level of transformations(mostly parsing,flattening of the fixed with files) and placing them in Delta Lake partitioned by dates. And this layer is our Raw layer.[7 days retention of data in LZ but in Raw no data will be deleted, it will be partitioned with date]

**3)Hammer solution zone** UsingADF,we are picking the data based on date partitions and copying it down to our DWH model.

Entire DWH capabilities/business logic (SCD Type1, Type 2, Facts,Dimensions, Surrogate key, Snowflake, Star) are kept inside Synapse.

We are following the traditional 3 layer data warehouse model, maintaining 2 different Dedicated SQL pools.

**Pool1: TO CATER ETL CAPABILITIES**

**MRR**: RR, [Truncate load table which will take only delta data/incremental data based on the date of last execution which we have kept in a control table and it will hold single days of data] ADF will load the data in MRR.

**STG:** RR, Entire DWH capabilities/business logic (SCD Type1, Type 2, Facts,Dimensions,identifying deltas, Surrogate key) resides here. Data will be loaded in STG from MRR using Synapse Data Flows.

**Pool2: TO CATER DWH CAPABILITIES: querying will be more**

**TMP:** RR, only for dimensions we are collecting the delta. UPSERT to update/insert data into the DWH tables.

**DWH:** Hash for Facts/bigger dimensions

Replicated for Dimensions: UPSERT/Merge to update/insert data into the DWH tables based on the delta on TMP

Once our data processing & storage are done, we expose Synapse to PowerBI using

2 layers of views.

**DWH views:** IT people(with select privilege): column level security

**USR views:** dedicated for Power BI dashboard, row level security

PowerBI

That's pretty much in a nutshell what we are doing in our project.

CICD pipeline for deployment(ADF and ASA)

2 fact tables, 14 dimensions, Snowflake since some of the dimensions are joining with them[Product\_consolidtaed]

Online Sale detail: contains all product

Online Order: contains dims like calendar, store, pickuppoint, pickuppointtype

History load of 1 year data with iterations of 12, monthly basis

Daily POSLOG→ approx 400MB

1-1.5 GB daily, yearly → 2000 million records, 2000 DWU units,

–Monthly approx 80million facts,225GB, yearly 960 million,, approx 3 TB

Files: POSLOG, NASA(product master), FILD(store master), FOFF, DATE, ECDT(product for online specific)

**1)Performance tuning:**

a)As far as data load is concerned, its taking 30 mins(adf→ 7-8 mins and Synapse 15-20 mins for 4 million fact data everyday)

→Hash for Facts/bigger dimensions; Replicated for Dimensions, speed of UPSERT is faster than Merge.

Breaking Data Flows(Check transformation logic: if possible break it), increasing DWU, Azure IR, no of cores/worker nodes, distinct queries.

b)As far as Power BI is concerned, we improved performance 20-22 secs to 6-7 secs with 500 DWU

570 million facts data

Reduce number of slicers in Power BI: Product\_consolidated table contains denormalized view of all the product\_id ie. NASA data, basically we are Snowflaking, uses a consolidated table in PBI instead of joining all the product hierarchy dimensions.

Synapse DWU,no of drivers and worker nodes, no of cores

Gathering stats of db, removal of orphan users

Automate scale up and scale down the DWU units;

Join with SK between facts and dimensions

In dimensions, change Business key to int datatype

Materialized view

**2)Data reconciliation:** for the fact table, product\_id is there but some of the product id are missing in product dimension. We were unable to tag a sale with a product

We assume product\_id to be similar to another field(product\_id=Nasa no).

Inn a file ECDT, Supplier\_item\_no = NASA no and Item\_no = Product\_id

So ECDT file is the linkage between nasa no and product\_id

There are some product\_ids which are not there in the online sales, but those are present in the product dimension,

We are filtering all such non-matching product\_ids and updating such records in product consolidated to “00 Unknown” for the descriptions and ‘-1’ for the id columns.

3)For dimensions, if Replicated then we can’t use Merge scripts, had to convert it into UPSERT logic.