
Simone Zambetti

Final project work

Agenda

- Project overview & obiettivi
- Big data analytics:
 - Background
 - Architettura
 - Visualizzazione – PowerBI & Data model
 - Row level access
- Advanced analytics (time series):
 - Background
 - Architettura
 - Risultati

Project overview

Obiettivi

Big data analytics

Target:

- Cruscotto per il monitoraggio CPM per il business “Eletrification” in ABB
- Ordinato e fatturato non erano tracciati a livello globale; nonostante ci fossero due diversi sistemi che aggregavano i dati da più di 40 ERP, non c'era un quadro completo. L'obiettivo della dashboard è di aggregare questi due sistemi combinando e armonizzando i diversi campi e riconciliandoli con il bilancio ufficiale.
- Migrazione fonte dati a warehouse aziendale in Snowflake, sostituendo una vecchia architettura che si affidava interamente a powerbi dataflows per l'ETL e l'estrazione tramite excel.

Advanced analitycs

- Creare un POC per forecast per l'ordinato di una singola product line usando deep learning e un modello lineare più semplice
- Connettere Databricks a snowflake e azure blob storage
- Visualizzazione del forecast per i prossimi sei mesi

Big data analytics

Dashboard

Migrazione

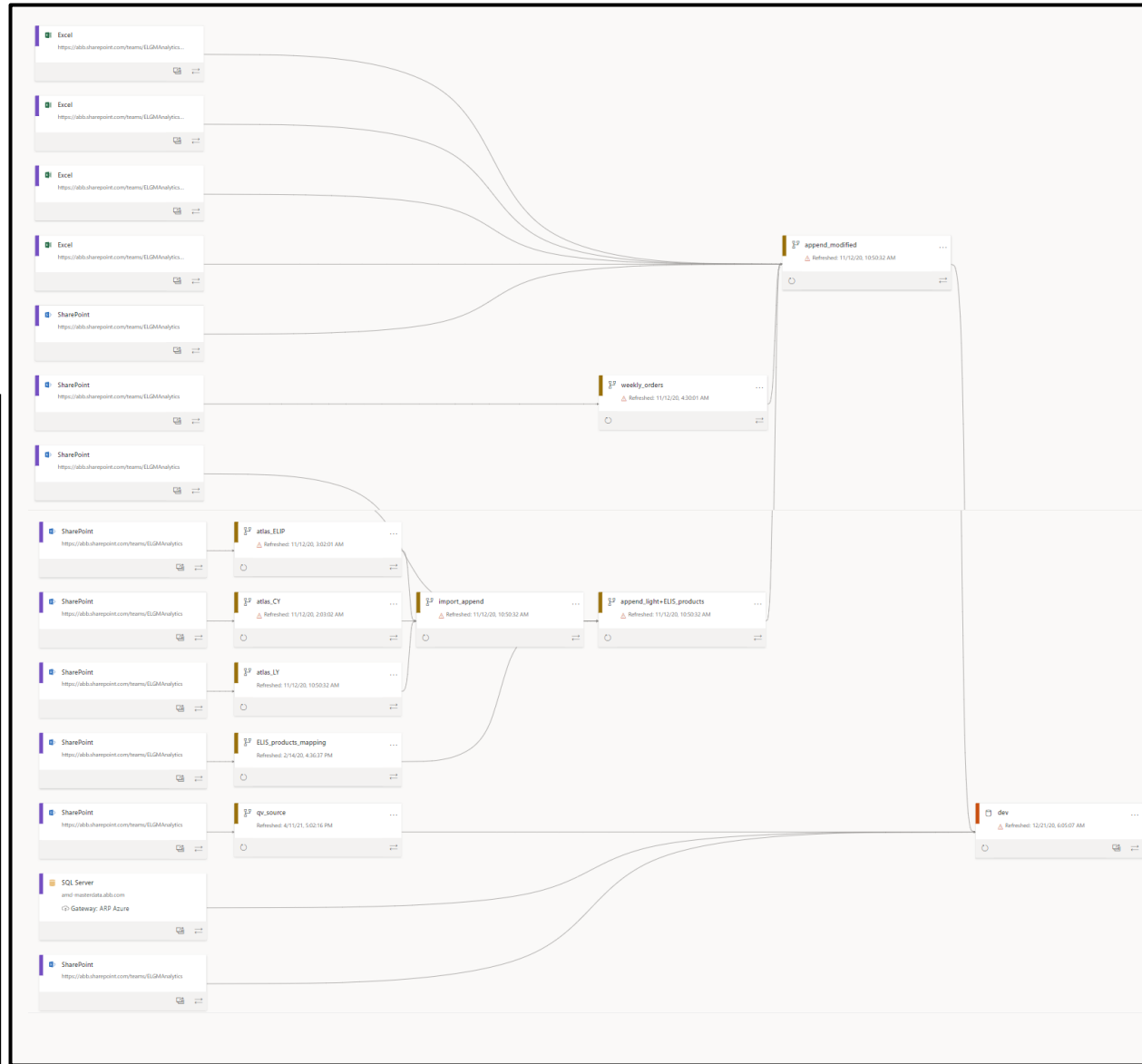
Da dataflows a snowflake

Background del progetto

In precedenza, vi era già una dashboard in produzione, la cui architettura si affidava a powerbi dataflows come transformation and load (a destra), inoltre che ad una workstation per effettuare estrazione tramite macro in excel. Questa architettura aveva molte lacune, quindi ho proposto una migrazione.

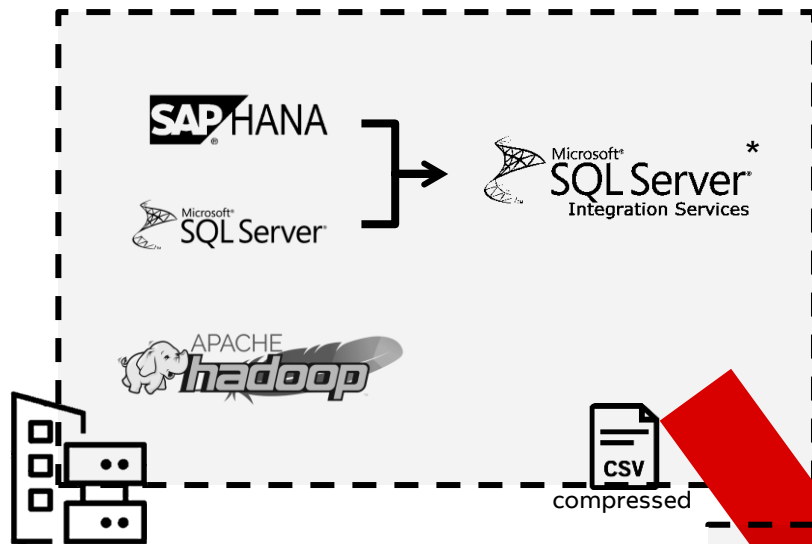
Il nuovo cruscotto powerbi è collegato a snowflake, che pesca da SAP HANA e Hadoop, rispettivamente collegati usando un job di SSIS che genera un csv zippato che viene poi spinto su azure blob. Da azure blob, c'è un ulteriore task snowflake che cancella la change table in staging e poi copia i dati dal blob su staging.

Dunque, un job di qlik compose inserisce/aggiorna i dati nella tabella di consolidation, da cui sono create due viste poi lette nel dashboard powerbi. Per Hadoop il processo è simile, ma il csv non viene generato tramite ssis bensì tramite un edge node di hadoop. Questa nuova architettura è meglio spiegata nella prossima slide.



Architettura

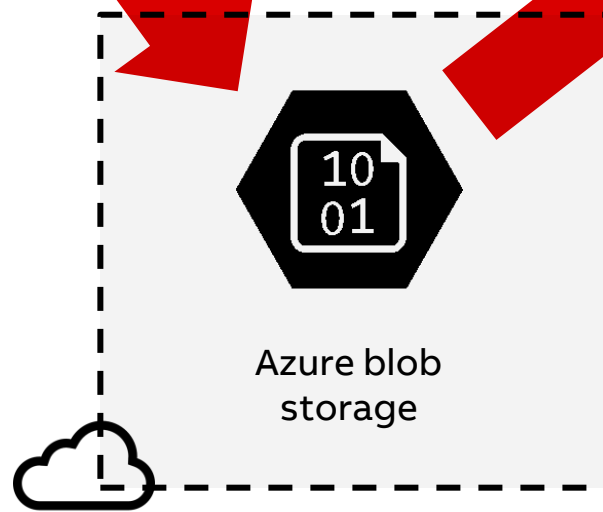
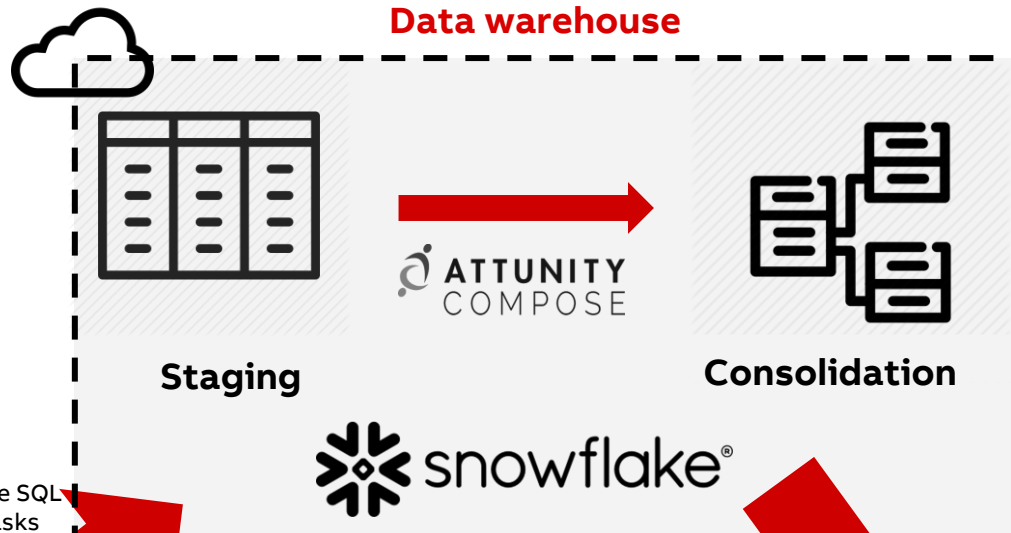
ABB on-premises



CSV
compressed

Snowflake SQL
«copy» tasks

Data warehouse



Data lake



PowerBI

Visualisation and consumption

*pacchetto SSIS realizzato da sviluppatori di ABB

Qlik compose

Qlik compose + snowflake

Databases

Shares

Data Marketplace

Warehouses

Worksheets

History

Abacus.ai POC

YTD update

New Worksheet

Find database objects

Starting with...

DEV_COMMON_DATA_VIEWS

DEV_CONSOLIDATION

DEV_DATAMART

DEV_HUB_CONSOLIDATION_REP_AMA

DEV_HUB_CONSOLIDATION_REP_AMC

DEV_HUB_CONSOLIDATION_REP_EU

DEV_PROJECT

DEV_STAGING

AMSAP

DWH_ATLAS

Tables

ACCOUNTS_RECEIVABLE

ORDERS_REVENUE

ORDERS_REVENUE_CT

Views

V_ACCOUNTS_RECEIVABLE

V_ORDERS_REVENUE_CT

DWH_DWP

DWH_ELIS

DWH_GIS

No Tables in this Schema

Views

V_MASTER

V_MASTER_AIM

V_MASTER_BPS_AIM

FIL_MANUAL

FIL_RBOOK

Tables

ABACUS_ORDERS_REVENUES_3RD...

ABACUS_REBATES

COUNTRY_MAPPING

1 select * from

2

3 with customer_table as (

4

Results

Data Preview

Table: DEV_STAGING.DWH_ATLAS.ORDERS_REVENUE

Data

Details

Filter result...

Row	OFISYEAR	OFISCVARNT	OCALDAY	OCURRENCY
1	2020	K4	20200309	EUR
2	2020	K4	20200317	CAD
3	2020	K4	20200421	SEK
4	2020	K4	20200430	EUR
5	2020	K4	20200430	CHF
6	2020	K4	20200320	USD
7	2020	K4	20200324	EUR
8	2020	K4	20200316	USD
9	2020	K4	20200407	USD
10	2020	K4	20200414	EUR
11	2020	K4	20200324	CNY
12	2020	K4	20200415	SEK
13	2020	K4	20200421	CAD
14	2020	K4	20200421	EUR
15	2020	K4	20200421	EUR
16	2020	K4	20200421	EUR
17	2020	K4	20200421	EUR

Table: ORDERS_REVENUE

Created on: 3/29/2021, 5:23:08 PM

Owner: RJIS_DEVELOPER

Rows: 13,973,845

Size: 1.2GB

Staging

Qlik

Compose for Data Warehouses

atunity.americas.ab.com | EUROPE/US/AM (Operator) | April 2020 (6.6.0.227) | Help

Edit Mapping - Map_ATLAS_ORDER_REVENUE

Filter... Data Profiler Data Quality Reset... Auto-Map Change View Null Updates...

Mapping Name

Map_ATLAS_ORDER_REVENUE

Landing Area Database

SR_ATLAS_STAGE_Landing

Schema

DWH_ATLAS

Source Type

Table View Query

View

V_ORDERS_REVENUE

3 Views

V_ORDERS_REVENUE

Landing Area Columns

OFISYEAR
OFISCVARNT
OCALDAY
OCURRENCY
MMICHANNL_MMICHNL2
MMICHANNL_MMICHNL2_T
MMINDSG_MMINDSG4
MMINDSG_MMINDSG4_T
MMINDSG_MMINDSG1
MMINDSG_T
MMICHANNL
MMICHANNL_T
MMISITEID
MMISITLC1
MMISITLC2
MMIAPPLIC
MMIAPPLIC_T
MPLPRD008_MPLPRD001
MPLPRD008_MPLPRD001_T
MPLPRD008_MPLPRD002
MPLPRD008_MPLPRD002_T
MPLPRD008

ATLAS_ORDER_REVENUE

Staging Columns

FISCAL_YEAR
FISCAL_VARIANT
CALENDAR_DAY
CURRENCY
MMINDSG_T
CHANNEL
SITE_LOCATION_ID
SITE_LOCATION_TOWN

Qlik Attunity Compose mapping

Abacus.ai POC

YTD update

New Worksheet

Find database objects

Starting with...

METADATA

O2C_NAT_NONERP

ORDER2CASH_NAT_NONERP

ORDER_TO_CASH_ELIP

ORDER_TO_CASH_NAT

ORDER_TO_CASH_RBOOK

Tables

DV_ABACUS_ORDERS_REVENUES...

DV_ABACUS_REBATES

DV_ATLAS_ACCOUNTS_RECEIVABLE

DV_ATLAS_ORDER_REVENUE

DV_COPA_CE10001

DV_COPA_CE1XXXX

DV_ELIP_DISTRIBUTOR_PART_MKT...

DV_ELIP_ORDER

DV_ELIP_SALES

DV_ELIS_ACCOUNTS_RECEIVABLE

DV_ELIS_SERVICE_ORDERS_REVEN...

1 select * from

2

3 with customer_table as (

4

Results

Data Preview

Table: DEV_CONSOLIDATION.ORDER_TO_CASH_RBOOK.DV_ATLAS_ORDER_R

Filter result...

Row	ID	FISCAL_VARIANT	CALENDAR_DAY	CURRENCY
1	13457348	K4	2020-03-23	CAD
2	2207538	K4	2020-03-27	USD
3	7936096	K4	2020-03-20	CNY
4	5348362	K4	2020-03-31	EUR
5	1699625	K4	2020-03-31	EUR
6	12563613	K4	2020-03-25	USD
7	5552558	K4	2020-03-03	RON
8	7611658	K4	2020-04-23	USD
9	5388402	K4	2020-03-31	CZK
10	10233911	K4	2020-04-10	EUR
11	1316665	K4	2020-03-16	EUR
12	3811729	K4	2020-04-30	CNY
13	4652618	K4	2020-04-17	SEK
14	8670285	K4	2020-04-28	EUR
15	11407153	K4	2020-04-20	USD
16	8472312	K4	2020-03-18	EUR
17				
18				
19				

DV_ATLAS_ORDER...

Preview Data

13,973,846 rows 1.5 GB

Cluster by

Columns

ID

FISCAL_VARIANT

CALENDAR_DAY

CURRENCY

CHANNEL

SITE_LOCATION_ID

SITE_LOCATION_TOWN

SITE_LOCATION_COUNTRY

APPLICATION

PRODUCT_ID

SERVICE_CATEGORY

Data Type

NUMBER(38,0)

VARCHAR(2)

DATE

VARCHAR(3)

VARCHAR(20)

VARCHAR(60)

VARCHAR(60)

VARCHAR(20)

VARCHAR(40)

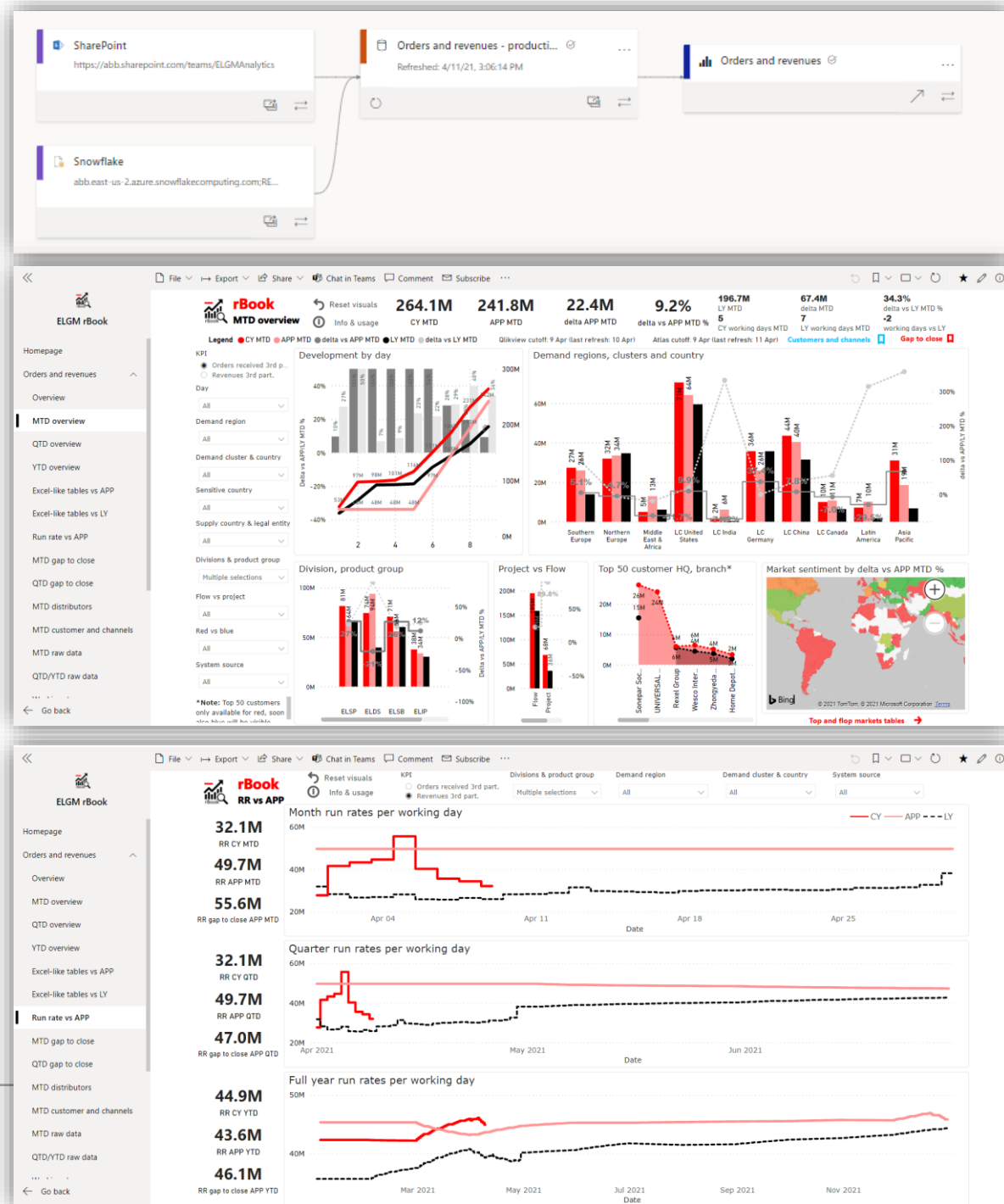
VARCHAR(3)

Consolidation

PowerBI

Visualizzazione e deployment in service

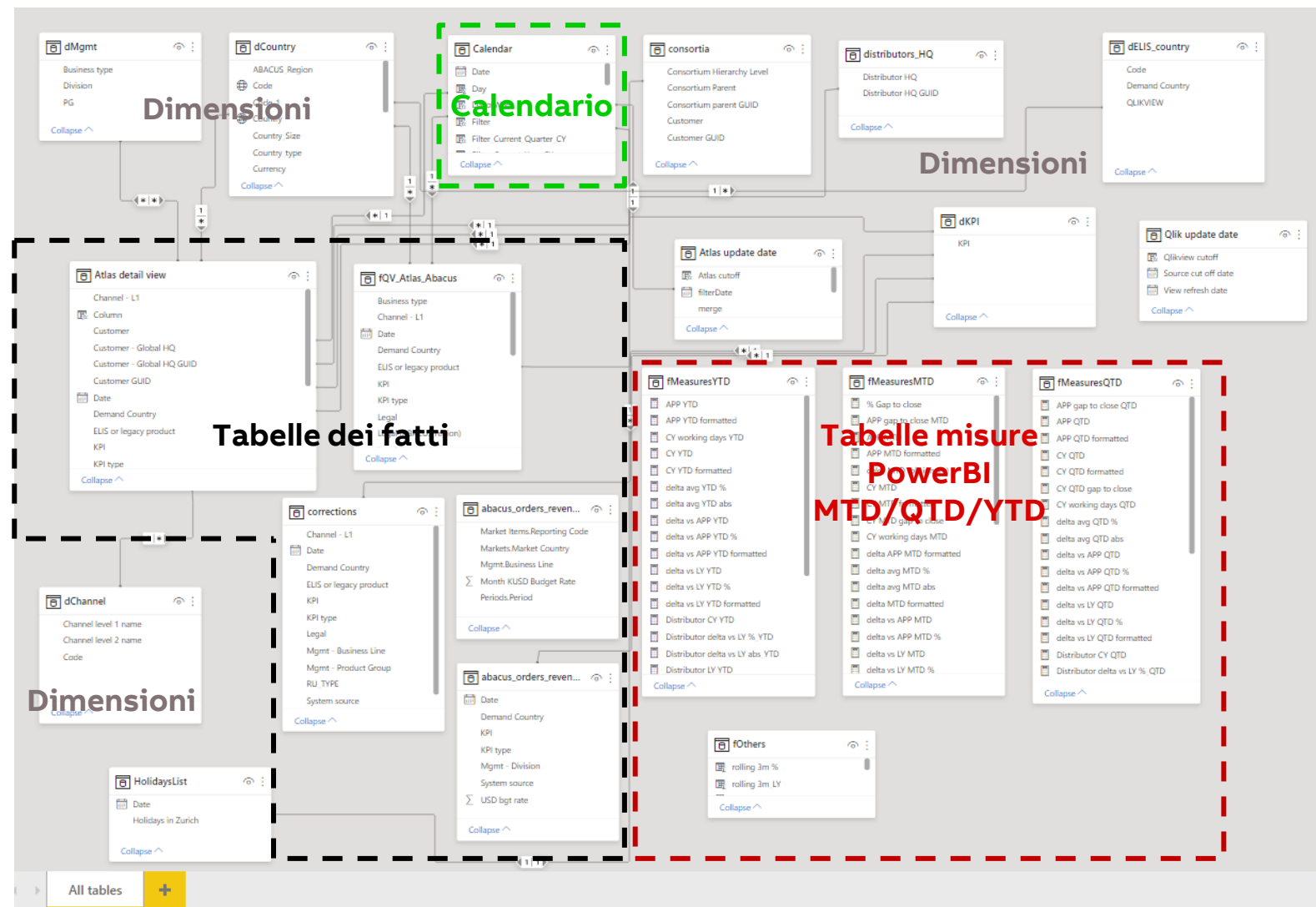
- Il layer di consolidation è stato dunque collegato a PowerBI insieme ad ulteriori mappature in Sharepoint. Il risultato su Powerbi service è un dataflow molto più semplice (primo screenshot in alto a destra).
- Per la visualizzazione, mi sono affidato a visual nativi di powerbi, per una migliore interattività. Ordinato e fatturato sono stati visualizzati secondo le seguenti principali dimensioni:
 - Business unit
 - Cliente
 - Regione e paese di appartenenza
 - Unità legale di vendita
 - Budget
- Numerose «measures» sono state create per effettuare benchmarking vs LY, vs budget o per visualizzare i trend quotidiani (screenshot in basso a destra).
- In service, la dashboard si rinfresca tutti i giorni alle 14 con un failsafe alle 15



Data model

Dashboard PowerBI

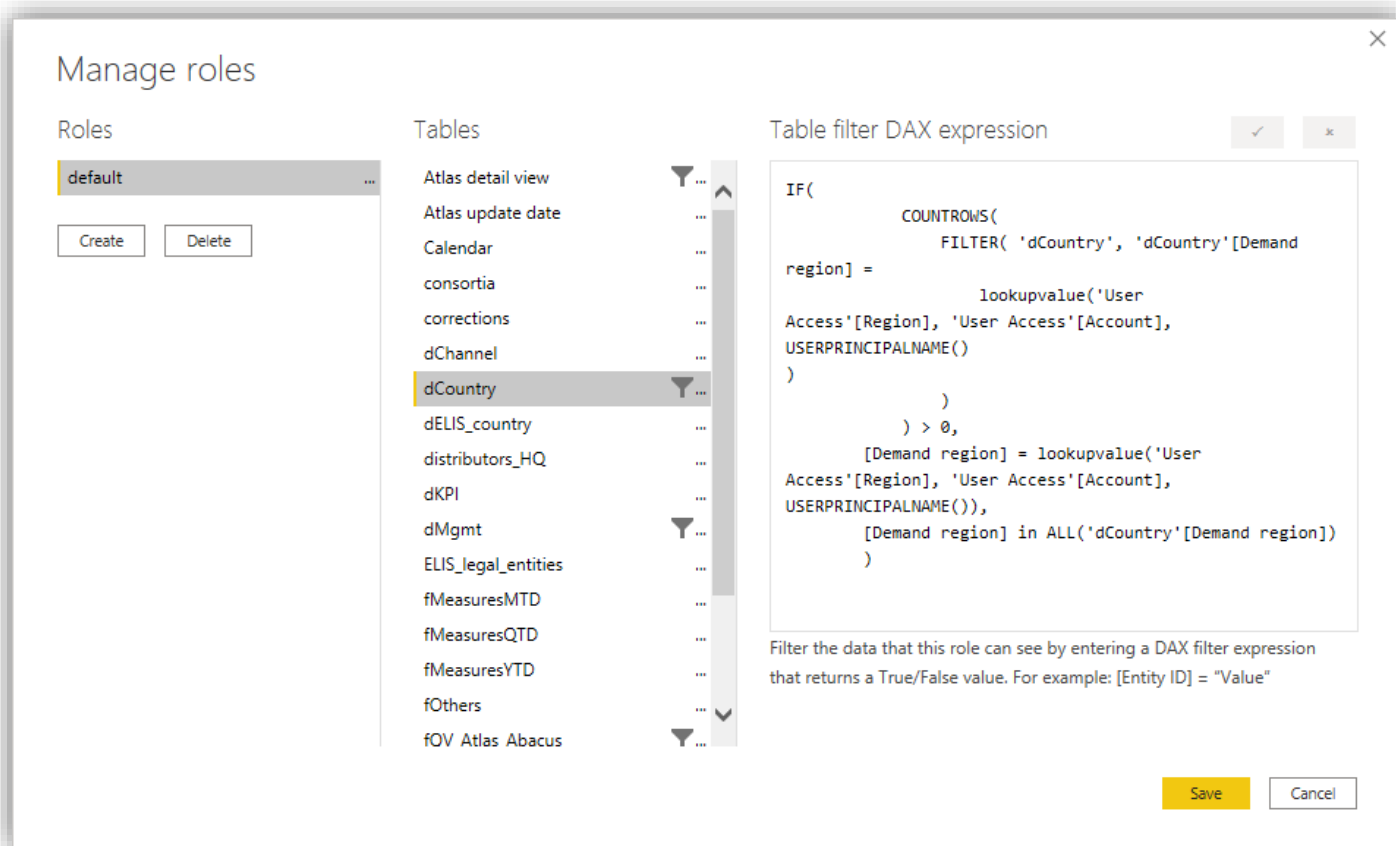
- A sinistra uno screenshot del servizio di analysis services di powerbi, ad esempio:
 - Dimensioni:
 - dKPI è la Dimensione: ad esempio, dimensione per selezionare ordinato/fatturato
 - dMgmt: dimensioni per business unit
 - Misure:
 - Benchmarking vs LY e Budget
 - Filtri per canale di vendite
 - MTD/QTD/YTD
 - Tabelle dei fatti: transazioni per ordinato / fatturato, più eventuali correzioni



Row level access

Limitazione accesso utenti

- La dashboard è stata rilasciata a circa 400 utenti
- Per limitare l'accesso per i singoli utenti a livello di business unit / regione, è stato creato un ruolo di default che filtra sulle tabelle di dimensione
- In questo modo, gli utenti non hanno accesso a regioni o business unit diverse a quelle di appartenenza.
- Alcuni utenti hanno comunque accesso a livello globale.

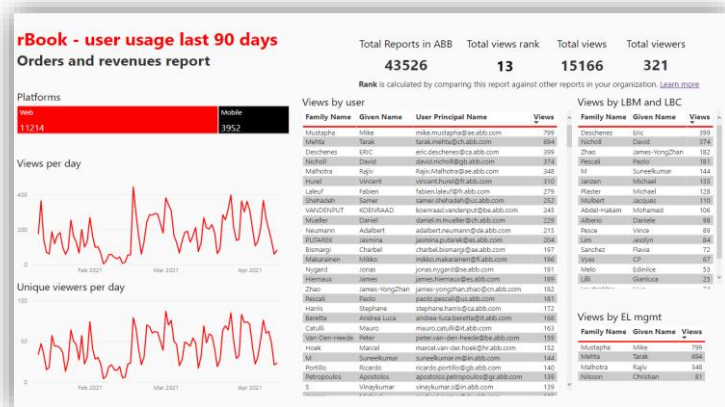


Conclusioni

Risultati, takeaways e next steps

Risultati e takeaways

- La dashboard è stata un successo, con 400+ utenti nella distribution list e di questi più di 300 sono attivi. Per questo una piccola dashboard per tracciare lo usage è stata creata in powerbi service* (come per screenshot sottostante). Al momento è la 13esima dashboard più utilizzata in ABB.
- Ci sono stati alcuni problemi con la stabilità del server di attunity e SSIS che poi sono stati stabilizzati con l'aiuto di sviluppatori senior



Next steps

- Non tutti i dati ingested in snowflake sono stati utilizzati per la realizzazione della dashboard.
 - Action: aumentare la mole di dati ingested da PowerBI utilizzando «incremental refresh» per visualizzare i dati fino a livello di singolo codice prodotto
- Migrare le «mappature» in sharepoint su blob e renderle disponibile in snowflake per altri progetti
- Il target è CPM per capire «cosa è successo e sta accadendo», in futuro si può prevedere l'integrazione nella stessa dashboard del forecasting POC, che vedremo nella prossima sezione



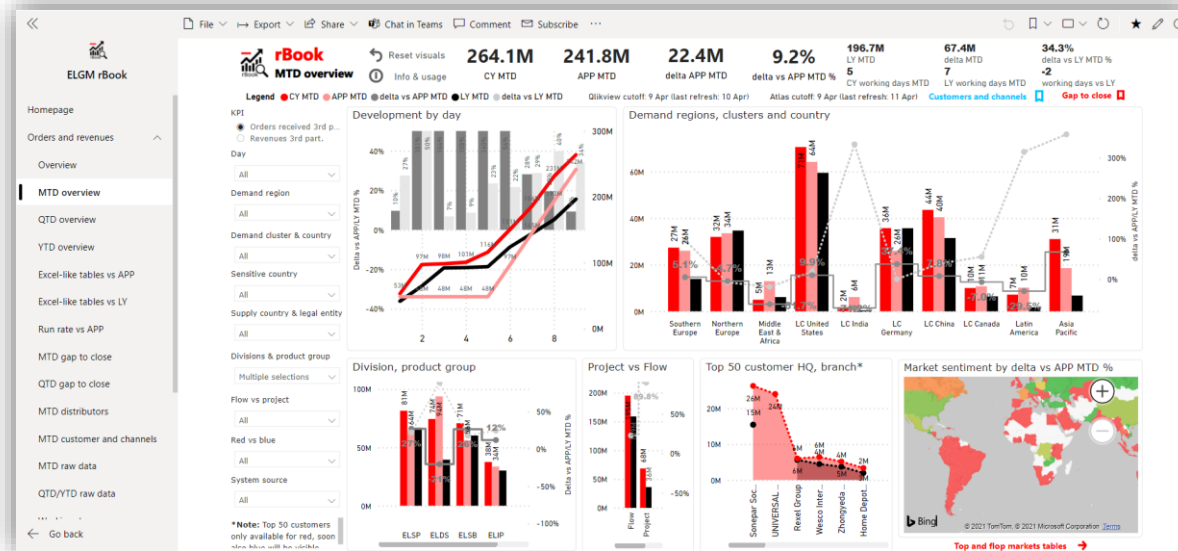
Advanced analytics

POC forecast

ABB Electrification business

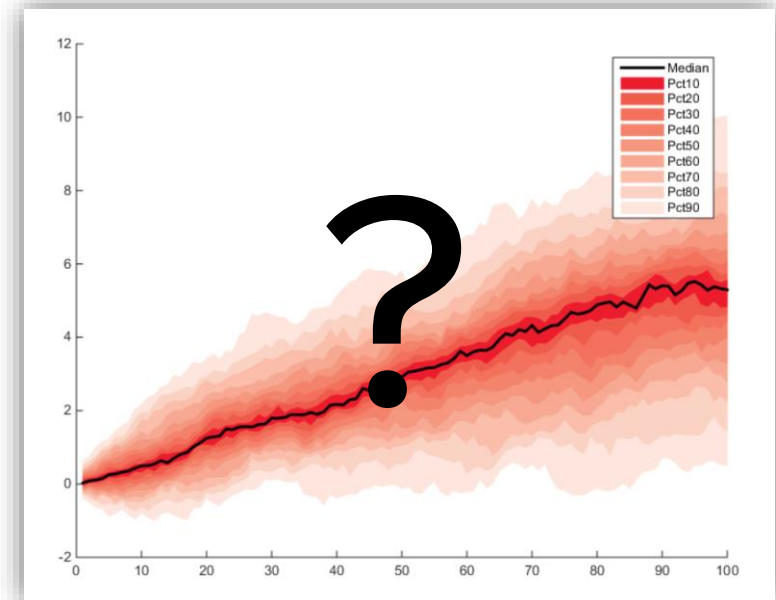
Decision making support per marketing and sales, finance

Reporting dashboard – presente e passato



La dashboard realizzata nella sezione precedente consente CPM per cosa è successo e cosa succederà

Predictive analytics – futuro

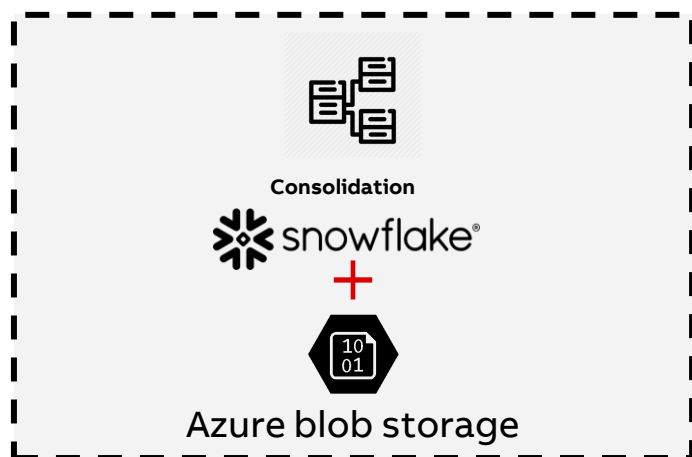


Al momento, per marketing e sales del business electrification, nessuno strumento di predictive analytics è in produzione

Forecasting POC

Architettura e dataset

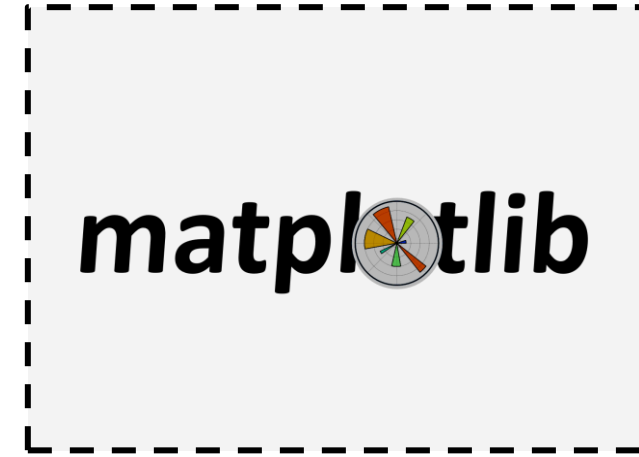
Sources



ML modelling + libraries



Visualisation

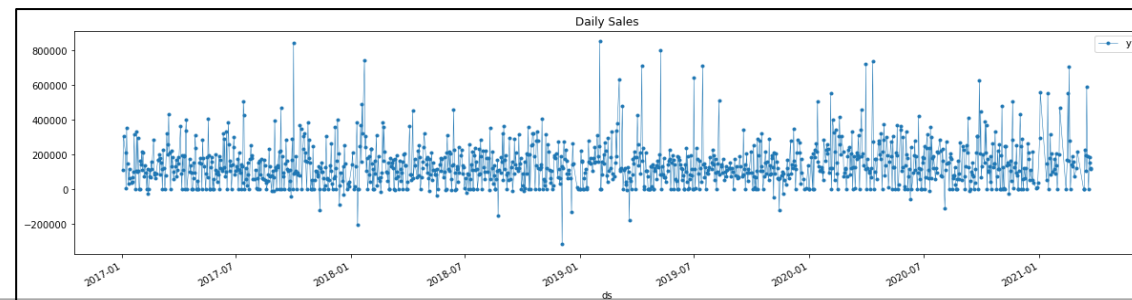


Architettura

Input – Serie storica dalle tabelle create in snowflake consolidation o se non presente, direttamente da blob storage. Il POC si limita alla serie storica di una product line

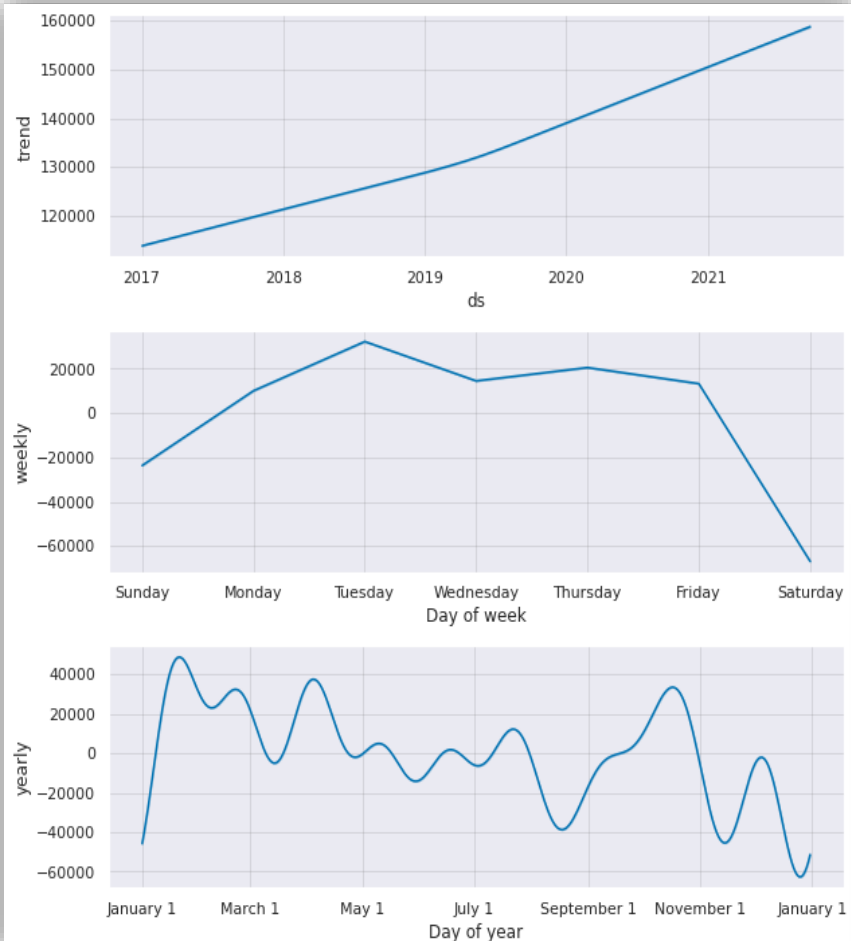
Modelling – databricks utilizzando pytorch e prophet

Output - Visualizzazione con matplotlib

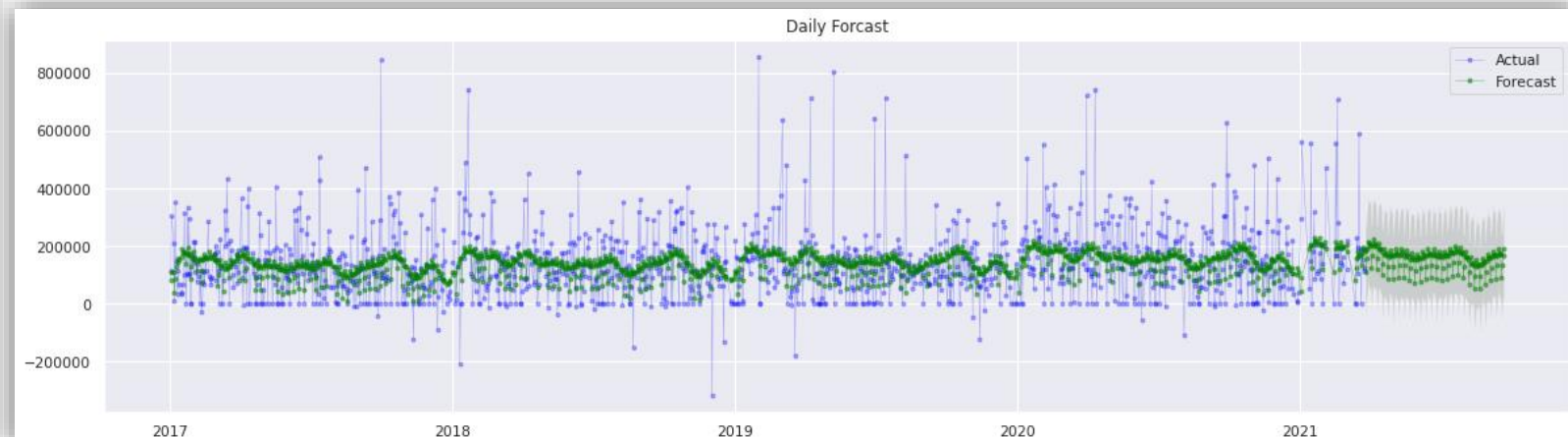


Facebook prophet

Timeseries analysis



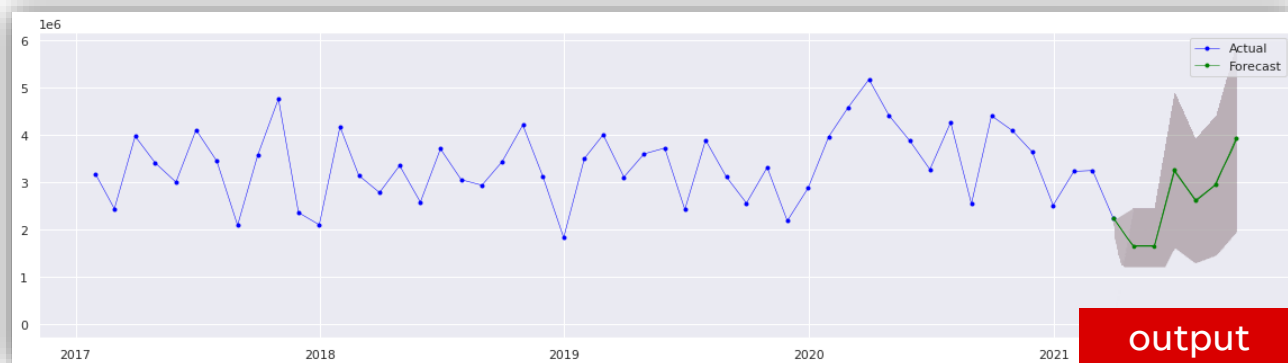
- Applicando Fbprophet sul dataset, si ottengono gli effetti di seasonality riportati a sinistra.
- Dunque, plottando la prediction per i prossimi sei mesi a livello giornaliero, come da grafico sottostante, si può osservare la prediction e lower/upper boundaries.
- Processando il dataset, sono stati rimossi gli outliers ma per la visualizzazione sono stati comunque inclusi.



Pytorch

LSTM for time series

- Per il POC, un'ulteriore obiettivo è stato applicare un modello di LSTM per il forecast della stessa time series
- L'input al modello è stato semplificato, dal momento che è stata usata una matrice mensile e lag 12 mesi anziché giornaliera, altrimenti sarebbe stata troppo complessa per il tempo a disposizione.
- L'errore sul test è di 80KUSD, che comparato ad una media di 3.2MUSD è accettabile.
- Tramite matplotlib, è stato plottato forecast* per i prossimi 6 mesi e actual



Microsoft Azure | Databricks

blob_connection (Python)

spark 3

Command took 0.02 seconds -- by simone.zambetti@it.abb.com at 4/12/2021, 12:23:50 PM on spark 3

Cmd 27

```
1 num_timesteps, df_train = create_ts_df(df_model_all['y'].values, start_index=0, end_index=None, history_length=12, step_size=1, target_step=1)
2 df_train
```

Out[23]:

	x_lag13	x_lag12	x_lag11	x_lag10	x_lag9	x_lag8	x_lag7	x_lag6	x_lag5	x_lag4	x_lag3	x_lag2	y
0	3168841	2437942	3972875	3408975	3002092	4107383	3446743	2089842	3567098	4764357	2358110	2099227	4173700
1	2437942	3972875	3408975	3002092	4107383	3446743	2089842	3567098	4764357	2358110	2099227	4173700	3144439
2	3972875	3408975	3002092	4107383	3446743	2089842	3567098	4764357	2358110	2099227	4173700	3144439	2781623
3	3408975	3002092	4107383	3446743	2089842	3567098	4764357	2358110	2099227	4173700	3144439	2781623	3353178
4	3002092	4107383	3446743	2089842	3567098	4764357	2358110	2099227	4173700	3144439	2781623	3353178	2578602
5	4107383	3446743	2089842	3567098	4764357	2358110	2099227	4173700	3144439	2781623	3353178	2578602	3706293
6	3446743	2089842	3567098	4764357	2358110	2099227	4173700	3144439	2781623	3353178	2578602	3706293	3053330
7	2089842	3567098	4764357	2358110	2099227	4173700	3144439	2781623	3353178	2578602	3706293	3053330	2937281
8	3567098	4764357	2358110	2099227	4173700	3144439	2781623	3353178	2578602	3706293	3053330	2937281	3433900
9	4764357	2358110	2099227	4173700	3144439	2781623	3353178	2578602	3706293	3053330	2937281	3433900	4211275
10	2358110	2099227	4173700	3144439	2781623	3353178	2578602	3706293	3053330	2937281	3433900	4211275	3118584
11	2099227	4173700	3144439	2781623	3353178	2578602	3706293	3053330	2937281	3433900	4211275	3118584	1827712
12	4173700	3144439	2781623	3353178	2578602	3706293	3053330	2937281	3433900	4211275	3118584	1827712	3503383
13	3144439	2781623	3353178	2578602	3706293	3053330	2937281	3433900	4211275	3118584	1827712	3503383	4005571
14	2781623	3353178	2578602	3706293	3053330	2937281	3433900	4211275	3118584	1827712	3503383	4005571	3109469
15	3353178	2578602	3706293	3053330	2937281	3433900	4211275	3118584	1827712	3503383	4005571	3109469	3599585
16	2578602	3706293	3053330	2937281	3433900	4211275	3118584	1827712	3503383	4005571	3109469	3599585	3722557
17	3706293	3053330	2937281	3433900	4211275	3118584	1827712	3503383	4005571	3109469	3599585	3722557	2423188
18	3053330	2937281	3433900	4211275	3118584	1827712	3503383	4005571	3109469	3599585	3722557	2423188	3891825
19	2937281	3433900	4211275	3118584	1827712	3503383	4005571	3109469	3599585	3722557	2423188	3891825	3113990

Input monthly matrix

```
1 from sklearn.metrics import mean_absolute_error
2 print('test mean absolute error: {}'.format(mean_absolute_error(y_act, y_pred)))
```

test mean absolute error: 78236.25

Command took 0.02 seconds -- by simone.zambetti@it.abb.com at 4/12/2021, 12:23:51 PM on spark 3

Error on test

Repo



<https://github.com/szambetti/BIBDA-final-project-public>