

Desenvolvido por: Guma Fernando Morais Pessoa

Formação: Gestão de T.I

Pós Graduação: Business analytics e big data (cursando)

Tema:

Utilização da plataforma Microsoft Azure para transferência de dados entre o banco de dados SQL e o ambiente de Big Data.

Descrição:

Criado um cenário fictício, baseado em *cloud computing* utilizando a plataforma Microsoft Azure. Onde teria um banco de dados contendo solicitações junto a Anatel referente às operadoras de telecomunicações. Neste cenário terá um banco de dados composto por mais de 5 milhões de observações, ao realizar consultas de agregações e entre outras mais complexas afim de gerar resultados para geração de *Dashboards*, foi notado que estava com uma alta latência para processar os dados, e seria necessário fazer a inserção desses dados para um ambiente que proporcionasse um melhor desempenho nas consultas, e então foi utilizado o recurso de Big Data Hdinsight da Azure para trabalhar com o cluster *Spark*, inserindo os dados em uma estrutura de processamento paralelo, e na sequência utilizar os resultados das operações para gerar *Dashboards* utilizando o Power Bi.

Atividades realizadas:

- Download das bases de dados referente às solicitações registradas na Anatel (dados abertos)
- Utilizado o Data Lake Analytics para unificar todos os arquivos
- Modelagem de dados (dimensional)
- Criação de um Banco de Dados Sql Server
- Criação de um processo de ETL utilizando o Data Flow do Azure Data Factory
- Criação de um ambiente de Big Data Hdinsight
- Utilização do sqoop para transferência dos dados
- Utilização do Hive e Spark Sql para processamento
- Utilização do Power Bi para criar visualizações

Observações:

Os dados utilizados foram coletados diretamente do site dados.gov.br, com o intuito de uso para o meu *aprendizado particular*.

Os resultados obtidos com as consultas **NÃO representam os valores oficiais**, pois a base pode conter valores faltantes e informações incompletas.

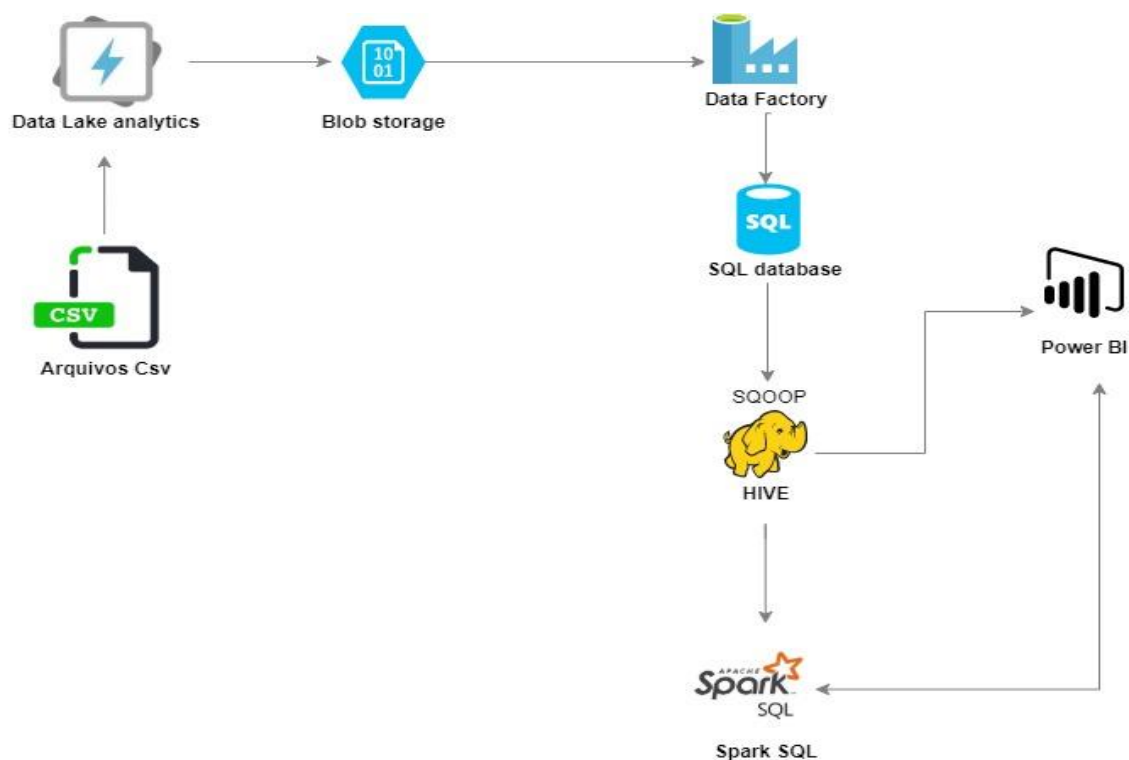
Último balanço que eu encontrei feito pela Anatel referente às solicitações está disponível em: <https://www.anatel.gov.br/institucional/mais-noticias/2497-anatel-divulga-balanco-dos-servicos-de-telecomunicacoes-de-2019>

O foco neste momento do desenvolvimento não está no pré-processamentos dos dados para gerar insights e sim na parte de levar os dados de um ambiente de banco de dados para um ambiente de big data, porém será gerado visualizações como forma de demonstrar os resultados obtidos.

O período analisado foi referente 2015 á 2019.

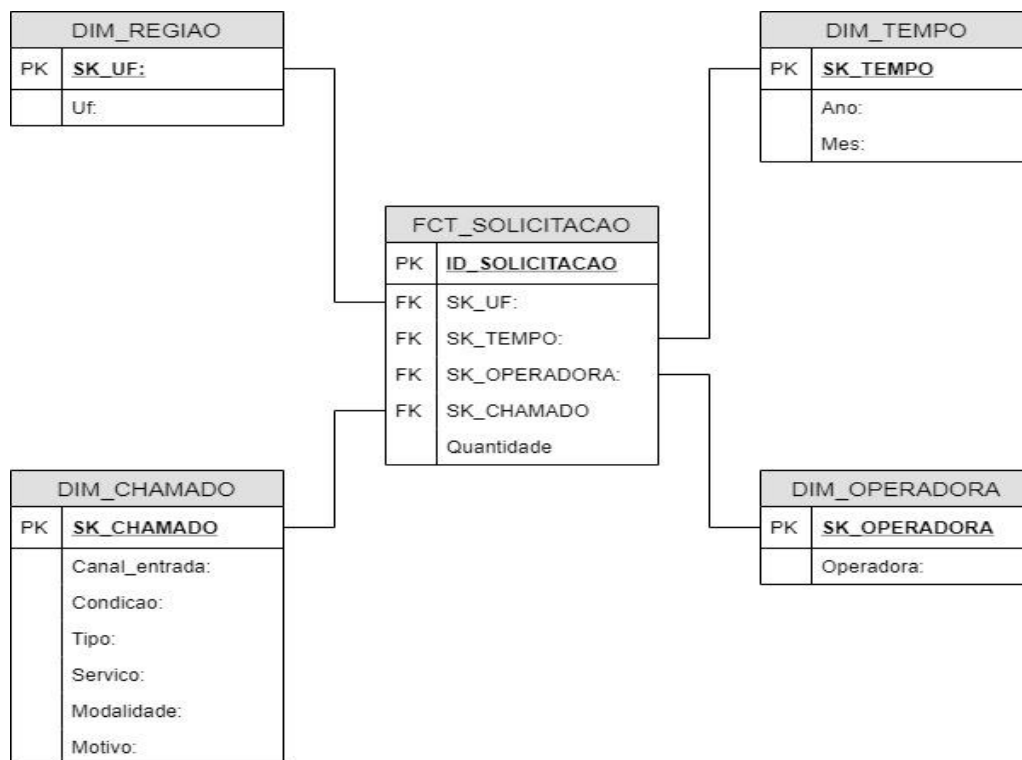
Base utilizada: <http://dados.gov.br/dataset/solicitacoesregistradasnaanatel>

Arquitetura



Etapas Realizadas:

Realizado a modelagem do banco de dados para criação das tabelas:



Criação dos bancos de dados

SQLQuery1.sql - gu...masql (hduser (65))* ↗ ✕

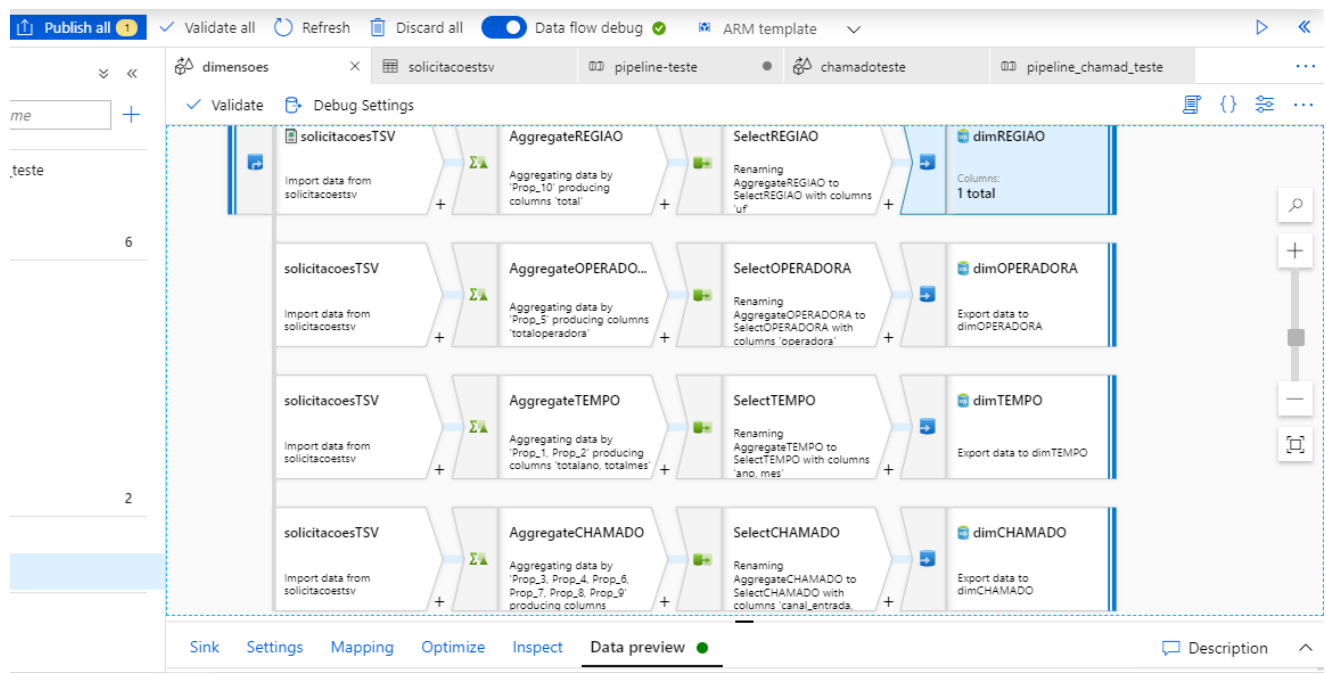
```
1 select * from information_schema.tables
2 |
```

90 %

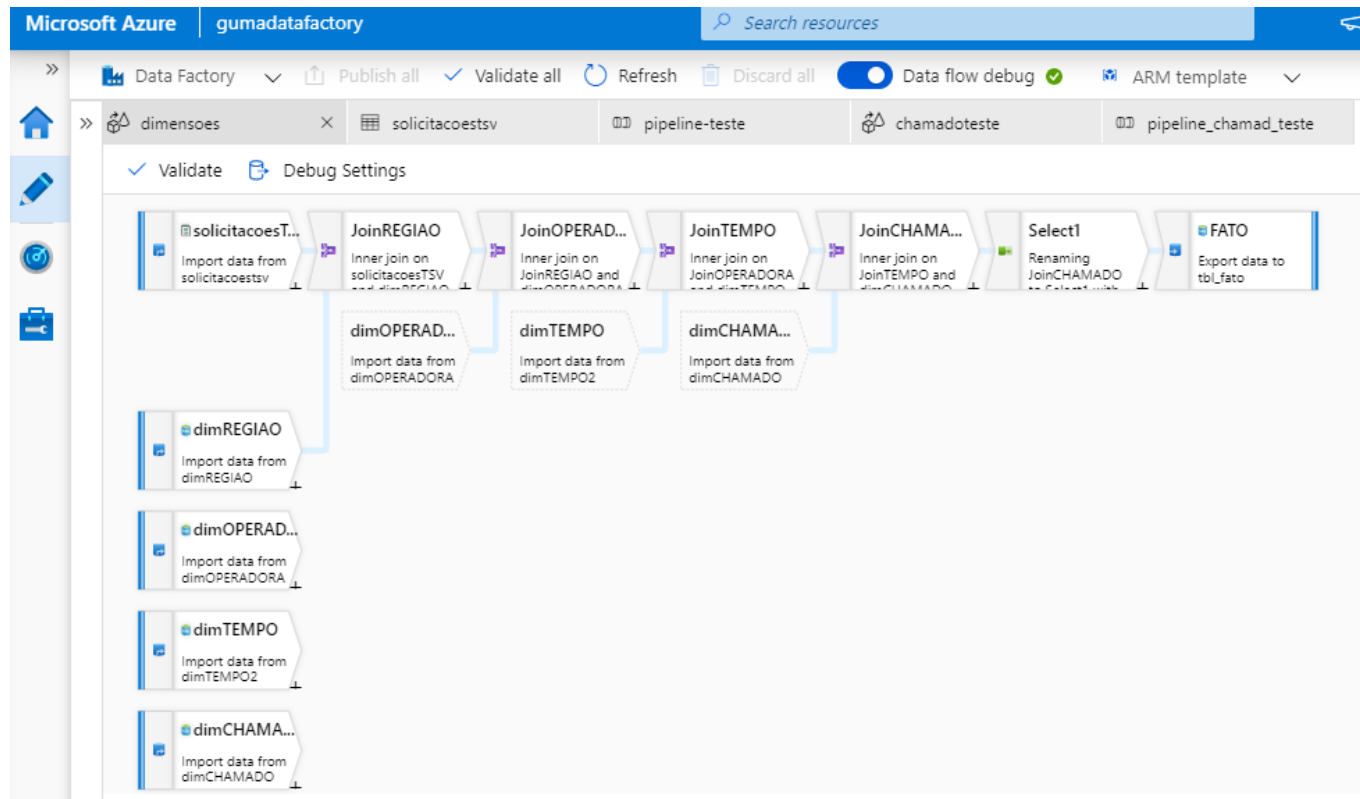
Resultados Mensagens

	TABLE_CATALOG	TABLE_SCHEMA	TABLE_NAME	TABLE_TYPE
1	gumasql	dbo	DIM_OPERADORA	BASE TABLE
2	gumasql	dbo	DIM_TEMPO	BASE TABLE
3	gumasql	dbo	DIM_CHAMADO	BASE TABLE
4	gumasql	dbo	DIM_REGIAO	BASE TABLE
5	gumasql	dbo	FCT_SOLICITACAO	BASE TABLE
6	gumasql	sys	database_firewall_rules	VIEW

ETL – Utilizado o data flow do Azure data factory , para realizar as cargas nas tabelas DIMENSÕES



ETL – Utilizado data flow do Azure data factory , para realizar a carga na tabela FATO



Consulta da tabela fato com os dados carregados

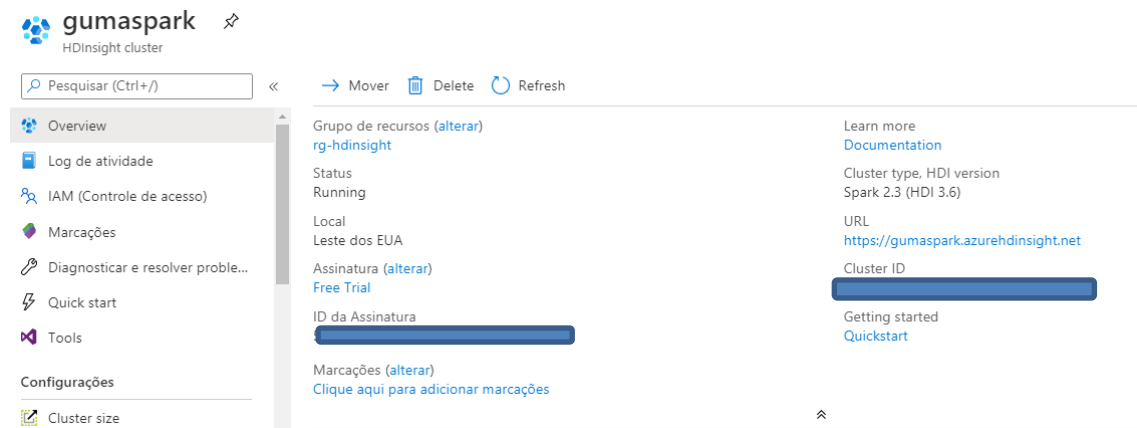
```
SQLQuery1.sql - gu...masql (hduser (65))* ✕  
1 select * from information_schema.tables  
2  
3  
4 select count(*) from [dbo].[FCT_SOLICITACAO]
```

90 %

Resultados Mensagens

	(Nenhum nome de coluna)
1	5164306

Criação de um cluster Hdinsight – Spark



The screenshot shows the Azure portal interface for an HDInsight cluster named 'gumaspark'. The left sidebar contains navigation links: Overview, Log de atividade, IAM (Controle de acesso), Marcações, Diagnosticar e resolver proble..., Quick start, Tools, Configurações, and Cluster size. The main content area displays cluster details: Grupo de recursos (alterar) 'rg-hdinsight', Status 'Running', Local 'Leste dos EUA', Assinatura (alterar) 'Free Trial', ID da Assinatura (redacted), and Marcações (alterar) with a link 'Clique aqui para adicionar marcações'. On the right, there are links for 'Learn more' and 'Documentation', and cluster specifications: 'Cluster type, HDI version' 'Spark 2.3 (HDI 3.6)', 'URL' 'https://gumaspark.azurehdinsight.net', and 'Cluster ID' (redacted). At the bottom right, there are links for 'Getting started' and 'Quickstart'.

Consultando as tabelas criadas no hive

```
0: jdbc:hive2://headnodehost:10001/> show tables;
+-----+-----+
|      tab_name      |
+-----+-----+
| hvdim_chamado      |
| hvdim_operadora    |
| hvdim_regiao       |
| hvdim_tempo        |
| hvfct_solicitacao  |
+-----+-----+
5 rows selected (0.59 seconds)
0: jdbc:hive2://headnodehost:10001/>
```

Utilizado o Sqoop para transferência entre o banco de dados e o CLUSTER

```
sshuser@hn0-gumasp:~$ sqoop import --connect $serverDbConnect \
> --table FCT_SOLICITACAO \
> --fields-terminated-by '\t' \
> --lines-terminated-by '\n' \
> --hive-database anatel \
> --hive-table hvfct_solicitacao \
> --hive-import -m 1
Warning: /usr/hdp/2.6.5.3025-2/accumulo does not exist! Accumulo imports will fa
il.
```

```

20/07/10 00:24:37 INFO mapreduce.ImportJobBase: Transferred 0 bytes in 68.7318 seconds (0 bytes/sec)
20/07/10 00:24:37 INFO mapreduce.ImportJobBase: Retrieved 5164306 records.
20/07/10 00:24:37 INFO mapreduce.ImportJobBase: Publishing Hive/Hcat import job data to Listeners
20/07/10 00:24:38 INFO manager.SqlManager: Executing SQL statement: SELECT t.* FROM [FCT_SOLICITACAO] AS t WHERE 1=0
20/07/10 00:24:39 INFO hive.HiveImport: Loading uploaded data into Hive

Logging initialized using configuration in jar:file:/usr/hdp/2.6.5.3025-2/hive/lib/hive-common-1.2.1000.2.6.5.3025-2.jar-
ve-log4j.properties
OK
Time taken: 2.861 seconds
Loading data to table anatel.hvfct_solicitacao
Table anatel.hvfct_solicitacao stats: [numFiles=1, numRows=0, totalSize=124994927, rawDataSize=0]
OK
Time taken: 5.076 seconds
sshuser@hn0-gumasp:~$

```

Verificando o resultado da importação, consulta do total de linhas na tabela fato

```

+-----+-----+
|   _c0   |
+-----+-----+
| 5164306 |
+-----+-----+
1 row selected (5.771 seconds)
0: jdbc:hive2://headnodehost:10001/>

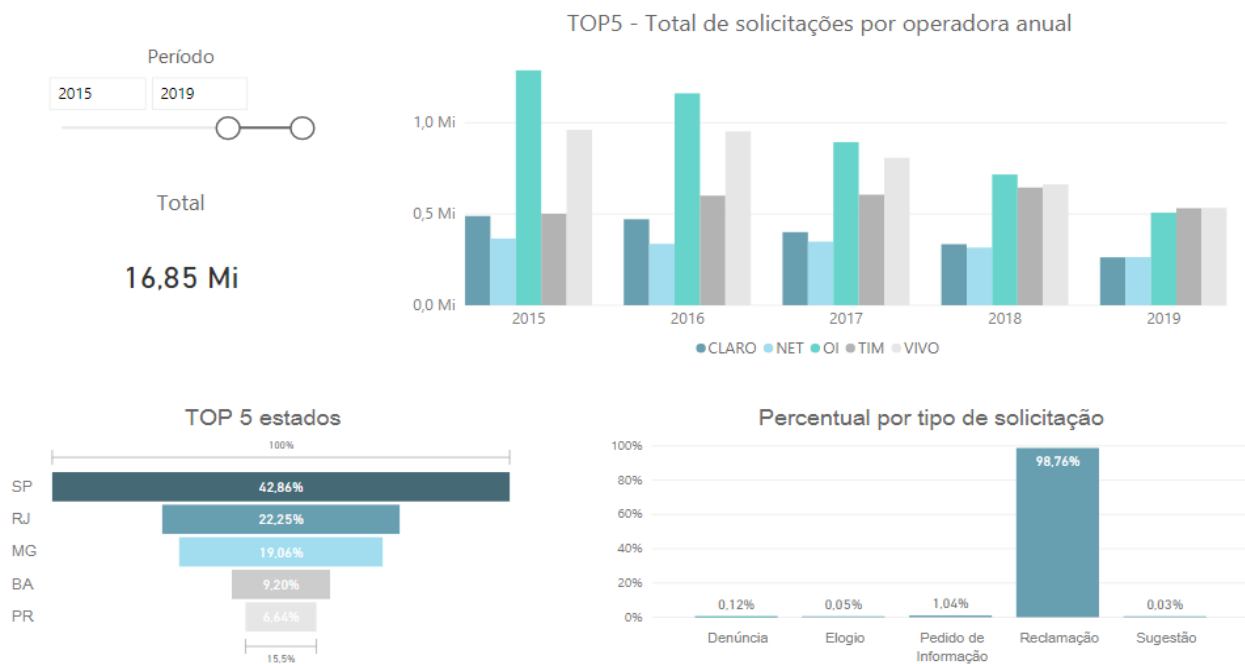
```

Utilização dos resultados gerados para gerar visualizações com o **POWER BI.**

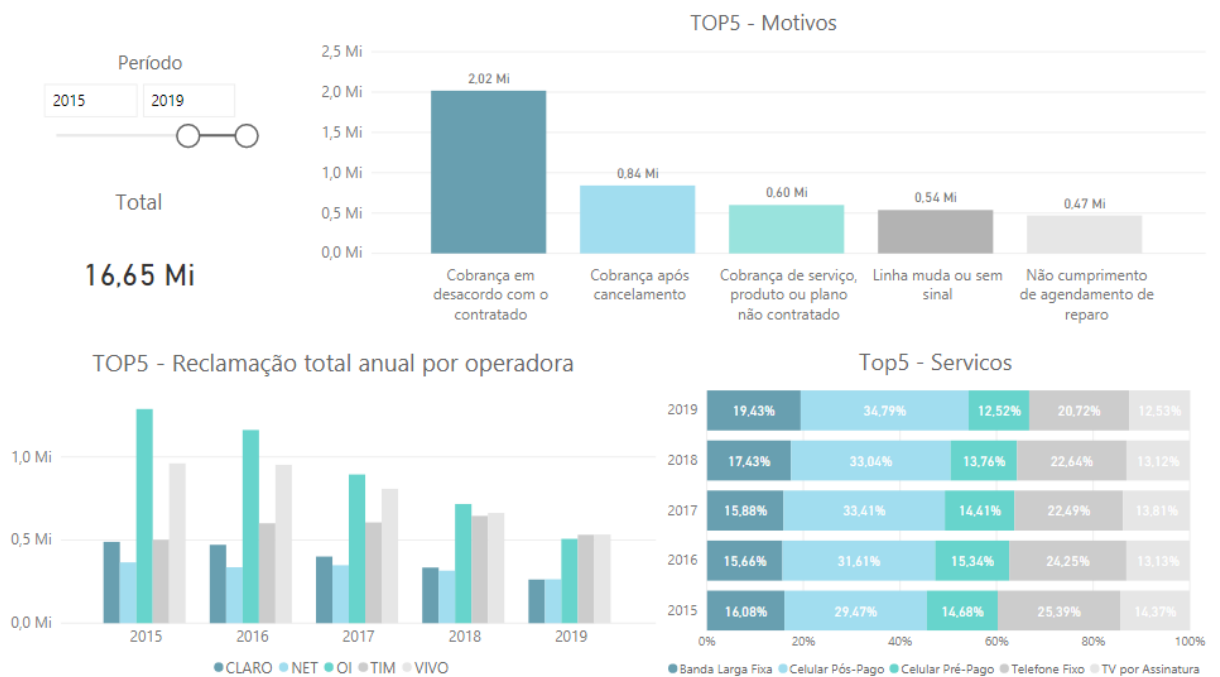
OBS: Para a visualização foi utilizada os dados referente o período de 2015 a 2020, devido a base não fornecer os dados referente alguns serviços.

Verificando ás cinco principais operadoras com maiores índices de solicitações em geral na Anatel.

- Verificando os cinco estados maiores índices de solicitações no período.
- Verificando o percentual dos tipos de solicitações referente ao período.

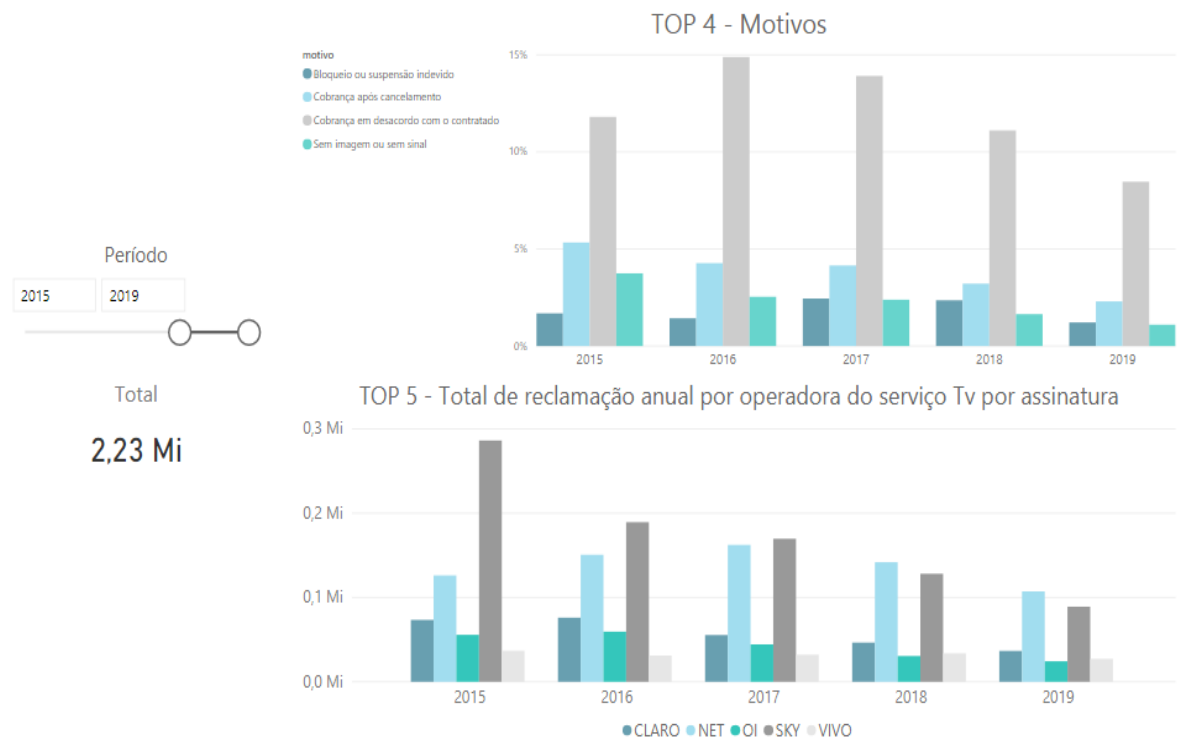


- Verificando os cinco principais motivos em geral.
- Verificando as cinco principais operadoras com maiores índices de **RECLAMAÇÕES**.
- Verificando os cinco principais **serviços** com maiores índices de reclamações



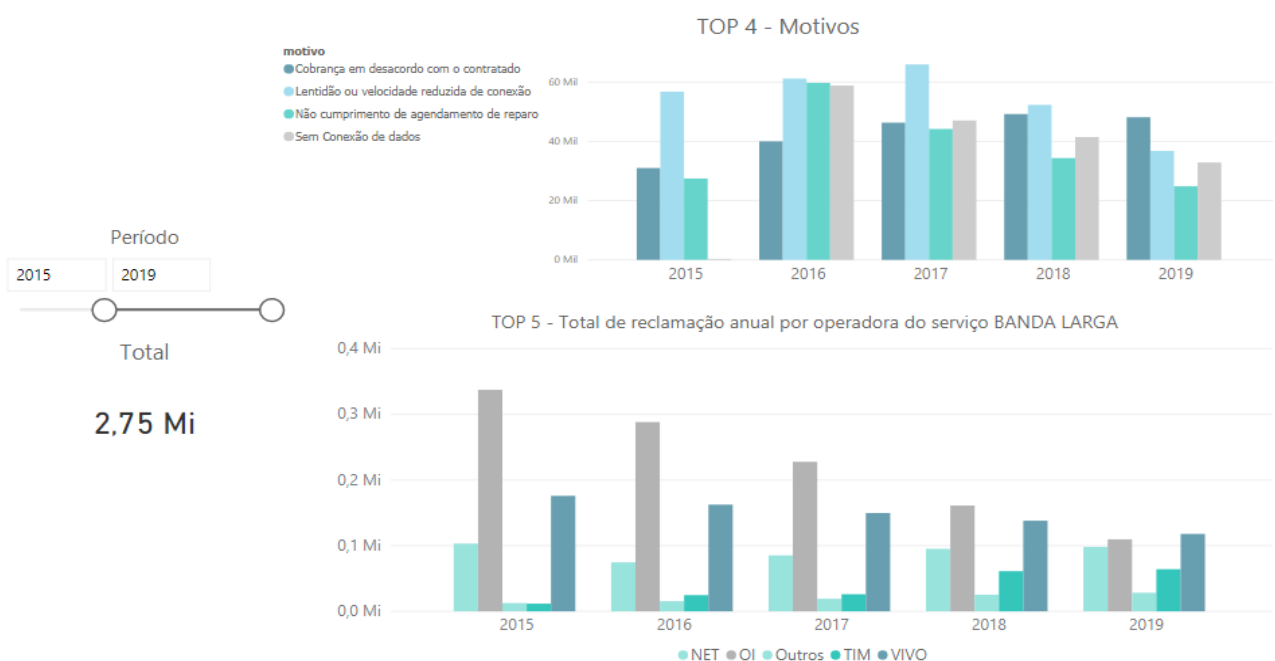
- Verificando as cinco operadoras do serviço **TV POR ASSINATURA** com maiores índices de reclamação referente ao período de 2015 á 2019.

- Verificando os quatros principais motivos das reclamações.



- Verificando as cinco operadoras do serviço **BANDA LARGA** com maiores índices de reclamação referente ao período de 2015 á 2019.

- Verificando os quatros principais motivos das reclamações.



- Verificando as cinco operadoras do serviço **TELEFONE FIXO** com maiores índices de reclamação referente ao período de 2015 á 2019.

- Verificando os quatros principais motivos das reclamações.



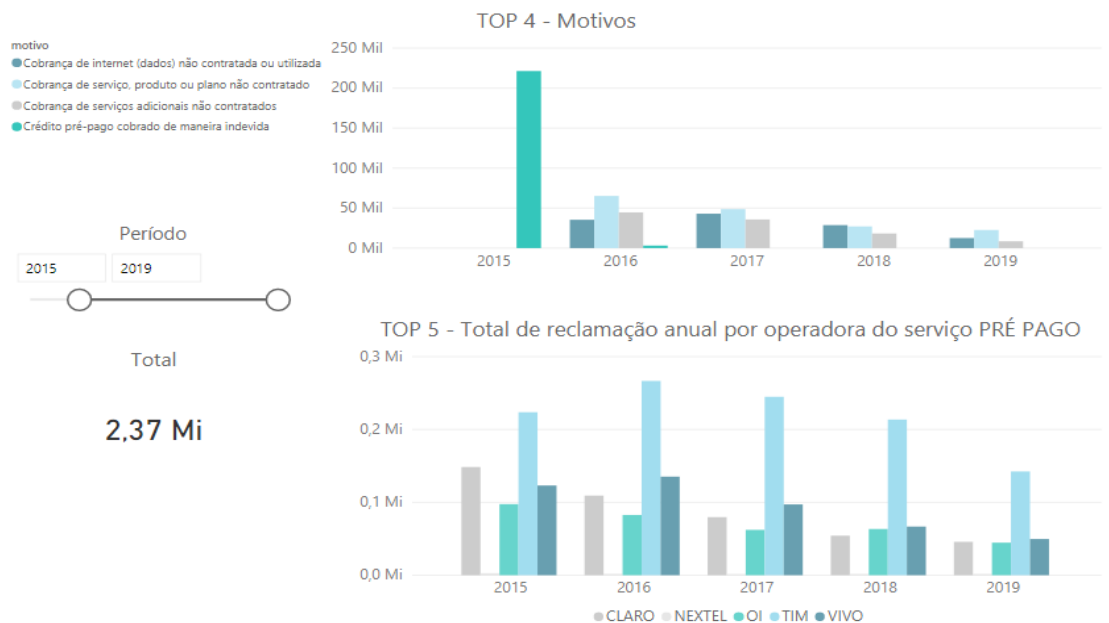
- Verificando as cinco operadoras do serviço **PÓS-PAGO** com maiores índices de reclamação referente ao período de 2015 á 2019.

- Verificando os quatros principais motivos das reclamações.



- Verificando as cinco operadoras do serviço **PRÉ-PAGO** com maiores índices de reclamação referente ao período de 2015 á 2019.

- Verificando os quatros principais motivos das reclamações.



Testes nos ambientes:

Query no banco de dados SQL com um tempo total de 01:55 minutos para executar.

```

15
16 =SELECT T.ano, O.operadora, C.servico, SUM (quantidade) as total FROM FCT_SOLICITACAO F
17 JOIN DIM_TEMPO T ON F.SK_TEMPO = T.SK_TEMPO
18 JOIN DIM_OPERADORA O ON F.SK_OPERADORA = O.SK_OPERADORA
19 JOIN DIM_CHAMADO C ON F.SK_CHAMADO = C.SK_CHAMADO
20 GROUP BY T.ano, O.operadora, C.servico ORDER BY total desc

```

ano	operadora	servico	total
2015	OI	Telefone Fixo	611693
2013	OI	Serviço Telefônico Fixo Comutado - STFC	559916
2016	OI	Telefone Fixo	550436
2014	OI	Serviço Telefônico Fixo Comutado - STFC	533110
2017	OI	Telefone Fixo	407811
2012	OI	Serviço Telefônico Fixo Comutado - STFC	391763
2013	OI	Móvel Pessoal	380282
2016	VIVO	Celular Pós-Pago	360326
2015	VIVO	Celular Pós-Pago	353384

consulta executada com êxito. | gumasql.database.windows.ne... | hduser (65) | gumasql | 00:01:55 | 1.103 linhas

Utilizado a mesma query no hive e com um tempo total de 00:38 segundos para executar.

```
0: jdbc:hive2://headnodehost:10001/> SELECT T.ano, O.operadora, C.servico, SUM (quantidade) as total FROM HVFCT_SOLICITACAO F
0: jdbc:hive2://headnodehost:10001/> JOIN HVDIM_TEMPO T ON F.SK_TEMPO = T.SK_TEMPO
0: jdbc:hive2://headnodehost:10001/> JOIN HVDIM_OPERADORA O ON F.SK_OPERADORA = O.SK_OPERADORA
0: jdbc:hive2://headnodehost:10001/> JOIN HVDIM_CHAMADO C ON F.SK_CHAMADO = C.SK_CHAMADO
0: jdbc:hive2://headnodehost:10001/> GROUP BY T.ano, O.operadora, C.servico ORDER BY total desc;
INFO : Session is already open
DEBUG : Adding local resource: scheme: "hdfs" host: "mycluster" port: -1 file: "/tmp/hive/hive/_tez_session_dir/5f67f220-f6e
f-430a-9657-1111390f6d5c/hive-hcatalog-core.jar"
INFO : Dag name: SELECT T.ano, O.operadora, C.servico,...desc(Stage-1)
DEBUG : DagInfo: {"context":"Hive","description":"SELECT T.ano, O.operadora, C.servico, SUM (quantidade) as total FROM HVFCT
_SOLICITACAO F\nJOIN HVDIM_TEMPO T ON F.SK_TEMPO = T.SK_TEMPO \nJOIN HVDIM_OPERADORA O ON F.SK_OPERADORA = O.SK_OPERADORA \n
JOIN HVDIM_CHAMADO C ON F.SK_CHAMADO = C.SK_CHAMADO \nGROUP BY T.ano, O.operadora, C.servico ORDER BY total desc"}
DEBUG : Setting Tez DAG access for queryId=hive_20200710011329_0b4fd0f8-2db5-4e07-82e9-8e417fcb7e8d with viewAclString=*, mo
difyStr=anonymous,hive
INFO : Setting tez.task.scale.memory.reserve-fraction to 0.30000001192092896
INFO : Status: Running (Executing on YARN cluster with App id application_1594338897186_0015)
```

t.ano	o.operadora	c.servico	total
2012	OI	Troncalizado (Trunking)	1
2012	SKY	Serviços da Anatel	1
2014	NET	Outros	1

,103 rows selected (38.324 seconds)
0: jdbc:hive2://headnodehost:10001/>

Discussão sobre o desenvolvimento e conclusões:

Primeiramente para mim foi um enorme desafio realizar essas etapas, devido ser o meu primeiro contato com a plataforma, e também ainda não ser da área de Dados e big data.

Todo o processo foi realizado na plataforma cloud Microsoft Azure, desde a unificação das bases extraídas através do Data Lake Analytics até o armazenamento no HIVE, e localmente foi utilizado o Power Bi para geração das visualizações.

Foi gerada uma query no banco de dados SQL e a mesma query no ambiente de big data, e foi verificado que o desempenho no ambiente de big data embora seja com uma quantidade de dados baixa com um pouco mais de 5 milhões de registros foi superior a 4x, ou seja, se o ambiente estiver bem planejado, para uma consulta com milhões de dados o desempenho será extremamente melhor em relação ao banco de dados transacional.

A partir das visualizações geradas, foi verificado que 98% das solicitações registradas na Anatel são de reclamações, sendo elas com maiores registros na região sudeste, e tendo como maior alvo de solicitação em geral a operadora Oi telecom. Após identificar que a reclamação tinha o maior índice, foi levantado os principais motivos durante ao período de 2015 á 2019, e constatado que a “cobrança em desacordo com o contratado” é o maior motivo entre as reclamações. Partindo deste princípio foi identificado os 5 principais serviços com maiores registros de reclamações e então tentar identificar quais os principais motivos durante o período da análise.

Diversos fatores podem influenciar a análise, como a quantidade de clientes ativos na carteira em um determinado serviço, ou cobertura de sinal em determinadas regiões.

Verificado que tanto nos serviços tv por assinatura, telefone fixo, pós-pago e no pré-pago, o maior motivo está em *relação a cobrança em desacordo com o contrato*, entretanto no serviço de Banda larga, o maior motivo está em relação a lentidão ou velocidade reduzida de conexão.

No serviço móvel verificado que a Tim ela é líder em reclamação nos últimos 3 anos 2017 e 2019, e no serviço de tv por assinatura a Sky é predominantemente a operadora com maior índice de reclamação durante o período.