T-Systems Int. GmbH, Telekom IT

Competence and Delivery Center Prague

CDCP

Big Data Use Cases in Telco Network

Author:
Ondrej Machacek

Responsible Manager:
Ondrej Benes

October 10, 2016

Abstract

Competence and Delivery Center Prague (CDCP) operates integration platform called iBMD (international Billing Mediation Device) which is responsible mainly for telco billing mediation [1] in four European countries (Germany, Czech Republic, Netherlands and Austria) under Deutsche Telekom group. iBMD collect charging data (CDRs [2]) from telecommunication network, processes and analyses them. In fact it means notification about every call, SMS/MMS or data usage of a T-Mobile customer across Europe goes through this system. iBMD collects binary data from network (so called CDRs-Call Detail Records), processes and sends them to different target systems, which are responsible i.e. for billing services, data warehousing, fraud detection, legal investigation and many more. IBMD deals with a lot of interesting, but also very sensitive and valuable data, five billion of CDRs is processed every day. Therefore, data security, correctness and completeness is a must in this area. With rise of data volume over the years, the need of data storage that would be able to handle and analyze all the data in- creased. Within the context of cost savings, open source Big Data project Apache Hadoop was chosen as a possible solution and subject for our interest and further investigation. The main motivation for this project is to take already existing knowledge of telco data, Hadoop and existing mediation system, and make it more intelligent. Our intention is to use Hadoop modules, such as Spark and its libraries for making predictions, clustering and machine learning. CDRs contain various fields reflecting customer behavior in the telecommunication network, these most interesting are: location, parties involved of in the service chain (who calls who for instance), identification of a data service accessed, timestamp, duration online, mobility etc.

1 Introduction

Overall DTAG footprint covers basically the whole Europe. CDCP (Competence and Delivery Center Prague) operates in the biggest markets in Europe and is responsible for storing and analyzing data reflecting customer actions. Introducing Hadoop should bring new ways of processing data and also new views on customers themselves. Following chapters describe several interesting use-cases which may be implemented on the data CDCP holds. CDCP owns two development hadoop clusters with pre-installed Cloudera hadoop distribution which are ready to be used for implementation and testing of the following topics. Please keep in mind data which iBMD holds are very sensitive in terms of customer data privacy. Therefore there are very restrictive conditions on any manipulation and distribution of these data. Generally data can't be shared without prior anonymization of important CDR fields.

How CDRs flow in the network is shown on the picture 1.

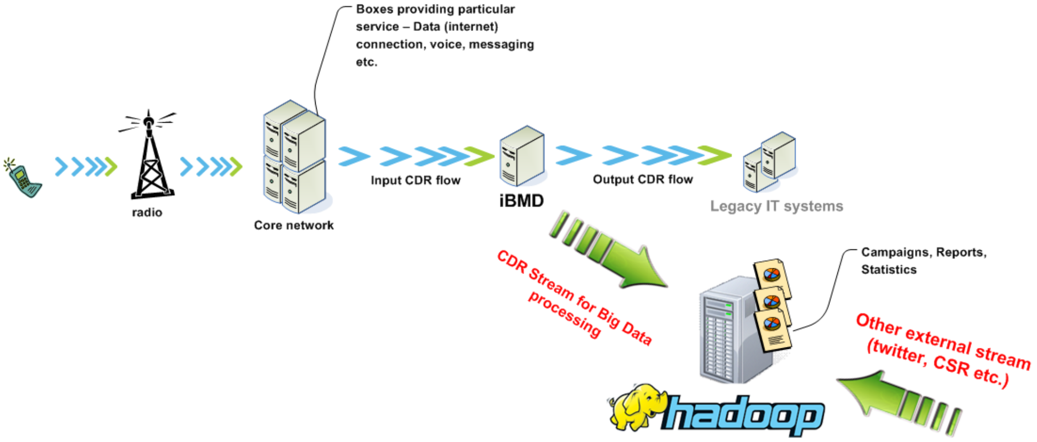
Boxes providing particular service – Data (internet) connection, voice, messaging etc.
radio
Core network
Input CDR flow
iBMD
Output CDR flow
Legacy IT systems
CDR Stream for Big Data processing
Campaigns, Reports, Statistics
Other external stream (twitter, CSR etc.)
hadoop

Figure 1: Telco Charging flow

2 List of Use-Cases

2.1 Active Archive

As mentioned above CDRs are archived to be ready for potential analysis. At the moment raw data as stored as a files on a SAN storage. This concept gives us very limited space for detailed analysis, is very slow and expensive. Active Archive shall be based on Hadoop and shall bring SQL like interface for accessing and querying CDRs. CDR is a structured data element which can be understood as a simple key, value set. Data are not flat, but can be organized in tree hierarchy (certain level of sub-structures) having up to hundred of fields. There is usually thousands of CDRs stored in single file, binary encoded. There is roughly tens of different CDR formats which are changing quite frequently. The aim of this project is to implement an SQL interface which can be used for searching in CDR files. It shall be dynamic in terms of data fields (dynamic column structure in where statement) and fast enough when searching in tens of billions of CDRs.

At the moment we have a prototype implemented. CDRs are stored in a Parquet format, SQL interface is managed by Cloudera Impala or Hive. This concept works well for static number of data fields, but cant be used when searching by a key which is not pre-configured. Tools used: HDFS, Talend as ETL, Spark for pre-processing and Impala as a SQL interface.

2.2 Location Insights

CDR data contain one very interesting information - location where customer operates. This can be used for various very interesting analysis especially considering how many CDRs iBMD processes and archives. Information about customer location is updated quite frequently and can be basically used for very precise tracking of a customer. When combined with CRM data we can have a very good starting point for analyzing certain demographic aspects such as:

- heat-maps reflecting gathering of people categorized by certain demographic criteria
- analysis for urbanistic studies - build city for people
- traffic analysis - predictions of traffic jams, potential traffic bottlenecks etc.

2.3 Household detection

Another topic where location data can be used. Based on identification of repeating activity, close-group detection and another socio-patterns it is possible to identify important spots for certain customers - home, sport club, work etc. These information can be used for again for certain demographic, but also commercial purposes.

2.4 Clustering, similarity/anomaly detection

CDR is a tree structured key-value data set. This fact gives an opportunity to understand it as a vector in N-dimensional space and apply clustering for certain very interesting analysis. Telco operator usually searches for:

- **similarity** - to identify most common behavior of customers to be able to introduce new service/tariff and attract as many people as possible
- **anomaly** - to search for potential fraud behavior (someone wants to use, but not to pay ;)).
- **smart categorization** - number of CDRs and generally amount of any other data generated by telecommunication network is rising quite rapidly so some smart pre-filtering is needed to be able to orient in this flood.
- **realtime** - all the points above will most certainly run in real time. This means models shall be recomputed on the fly, on streamed data.

Create models based on CDR data which help to fulfill these tasks shall be the aim of this topic. Clusters with pre-installed Hadoop, Spark and R are available already.

2.5 Data encryption and anonymization

As mentioned above, data produced by telecommunication network are very sensitive and can be accessed by any party without explicit permission. On the other hand information which can be gathered out of this data is very interesting for parties outside DTAG group (government authority for instance). Therefore it is very important to have the data encrypted on the HDFS, without affecting performance. And also it is necessary to consider some anonymization concept how to hide originating customer, but not to loose precision for potential further analysis - identification of demographic and social patterns. Very important fact here is that group of DTAG customers is in fact a limited set.

References

[1] 3GPP TS 32.240 *charging architecture*

[2] 3GPP TS 32.299 *Diameter charging applications*