Set Top Box Data Analysis

Introduction

The objective of the project is to analyze set top box data and generate insights. The data contains details about users' activities like tuning a channel, duration, browsing for videos, purchase video using VOD (video on demand), etc. Process the data using Spark and solve the problem statement listed in KPIs sections

Data Download Link

https://www.dropbox.com/s/awkr0jerj5sxx8a/Set_Top_Box_Data.txt?dl=0

Sample Data

```
11001^1^100^2015-06-05 22:35:21.543^<d><nv n="ExtStationID" v="Station/FYI Television, Inc./25102" /><nv n="MediaDesc" v="19b8f4c0-92ce-44a7-a403-df4ee413aca9" /><nv n="ChannelNumber" v="1366" /><nv n="Duration" v="24375" /><nv n="IsTunedToService" v="True" /><nv n="StreamSelection" v="FULLSCREEN_PRIMARY" /><nv n="ChannelType" v="LiveTVMediaChannel" /><nv n="TuneID" v="636007629215440000" /></d>^0122648d-4352-4eec-9327-effae0c34ef2^2016060601
```

Format of Posts Data

- Server-Unique-Id
- Request-Type
- Event-Id
- Timestamp
- XML with tags of name and value
- Device Id
- Secondary Timestamp

KPIs

1. Filter all the record with event_id=100
   i. Get the top five devices with maximum duration
   ii. Get the top five Channels with maximum duration
   iii. Total number of devices with ChannelType="LiveTVMediaChannel"
2. Filter all the record with event_id=101
   i. Get the total number of devices with PowerState="On/Off"
3. Filter all the record with Event 102/113
   i. Get the maximum price group by offer_id
4. Filter all the record with event_id=118
   i. Get the min and maximum duration
5. Filter all the record with Event 0
   i. Calculate how many junk records are thier having BadBlocks in xml column
6. Filter all the record with Event 107
   i. group all the ButtonName with thier device_ids
7. Filter all the record with Event 115/118
   i. Get the duration group by program_id
   ii. Total number of devices with frequency="Once"