**Purpose:**

To analyze and publish the NPI data as per the need of the customer to a dashboard.

**Scope:**

Scope of this POC is to consider the NPI data as input and transform the data with the required fields out of input data (NPPES Data Dissemination ) and transform them depending on the look up (NPPES Data Dissemination code values – Weekly updates and Monthly Deactivation Update, codes ) tables provided.

**Understanding /Assumptions:**

The source data at http://download.cms.gov/nppes/NPI\_Files.html needs to be considered as an input (sample: npidata\_20160411-20160417.csv).

1. Weekly and monthly deactivation data should be considered for transforming the data (inserting the new record or updating the deactivation date in the record.
2. Enrich the data fields to their descriptions depending on the codes in the data for the appropriate mapping fields.Codes are available here:



1. Fields to be enriched are to be provided as there are multiple items in the data value

**Architectural considerations:**

Depending on the initial understanding and discussions on May 11th 2016 it is a purely a batch processing majorly focusing on the enrichment of the data and process the input data to provide the processed data as a CSV file and which acts as an input to the front end. The scope of this POC is to enhance the input data (sample: npidata\_20160411-20160417.csv) depending on the weekly and monthly files which are available at at <http://download.cms.gov/nppes/NPI_Files.html>

The weekly data should merge with monthly data based on below two conditions.

1. If PK of weekly file matches with PK of monthly file we need to update the record of monthly file with weekly file record.
2. If PK of weekly file doesn’t match with PK of monthly file we need to insert the weekly file record as a new record in monthly file.

**Technical considerations:**

The data loads are huge **(Size TBD)** and they are flat files considering the Big Data Hadoop and its eco-system stack for this POC.

**Initial considerations:**

* + Hadoop **M/R (YARN) or Pig Latin** for data processing.
  + **Hive** for ensuring the data available to the external users through interface
  + **Flume/ shell** scripts for downloading the data to Hadoop cluster.
  + Cloudera Distribution for Hadoop **(CDH5.4)** is considered as most of the eco-systems (Hive, Pig and spark) are pre-installed and configured additional configuration from the developers towards infrastructure is not required saves lot of development time.
  + User interface from the back end is provided through a **java utility**.
  + The POC can also be implemented through Pig for the extraction of the data and provide the user interface.

**High Level Design:**



**Benefits of this approach:**

1. The extraction job is done in a batch processing mode and is available all the time. There will not be a frequent runs of map reduce for every request from the front end.
2. Front end application will pass the parameters (NPID or any other search criteria) Java utility will transform the parameters and generate a query which will trigger on hive external table.
3. Considering Hive partition to ensure the fields are correctly displayed to front end and reduces the time of mapping at the front end.
4. As extracted data is already available and partitioned by field **(TBD)** the hive query just hits that partition and data retrieval will be faster.

**Limitations:**

1. Map Reduce uses lot of Disk level processing which may lead to disk crash ( Basic disadvantage of Hadoop nomenclature) but this limitation can be addressed by replication factor.
2. Processing time of Hadoop Map Reduce programs are slow compared to other eco-systems.
3. Development time for Map Reduce is more compared to other technologies.
4. For effective batch processing MapReduce is considered alternatively Pig is preferred option.
5. For Integration with other stacks Map Reduce based code requires extra learning curve.

**Note:**

* For this POC we considered this as a unique piece, integration with other components is **NOT** taken in to account.
* Encryption of the data and flow of the data in secured way are out of scope of this POC. Default security which comes from CDH will be used for the data transmission to the Hadoop single node cluster. POC will be demonstrated on this only.