Big Data Technologies

Agenda

- Spark ML
- HBase

Advanced Analytics

- Analysis vs Analytics
  - Analysis --> Past data -- Understading data, Summarizing data, Visualizing data.
    - Python: Pandas, Numpy, Matplotlib, etc.
    - R: Frames, Stats, Charts, etc.
    - Excel: Ordering, Filtering, Pivot table/charts, etc.
    - PowerBI: Charts, Reports, etc.
  - Analytics --> Future/Predictions
    - Machine Learning --> Single system --> Languages: Python, R, C++, etc
    - Machine Learning --> Distributed systems --> Mahaout on Hadoop (outdated), Spark ML, etc.

Analytics

- Analytics refer to various techniques to solve core problem of deriving insights and making predictions/recommendations.
- Most common tasks are:
  - Supervised Learning
    - most common
    - using historical data train model
      - data have label (dependent variable)
      - data have features
    - Includes
      - classification: predict disease, predict purchase or not, classify images, ...
      - regression: predict sales, predict viewer count, ...

SUNBEAM INFOTECH

- Recommendation Engines
  - movie recommendation, product recommendation, ...
- Unsupervised Learning
  - find pattern or discover underlying struct in data
  - clustering, anomaly detection, topic modeling, ...
- Graph Analytics
  - based on graph data struct algos
  - fraud detection, classficiation, recommendation, find patterns in social network, ...

Analytics Process

1. Collect relevant data.
2. Clean & inspect the data -- EDA.
3. Feature Engg -- Extract features into numerical vectors
4. Build ML model using portion of data (training set).
5. Evaluate model using portion of data (test set).
6. Serve ML model to predict, recommend, ...

Spark ML

1. Collect relevant data.
   - Spark can get data from any source -- HDFS, S3, RDBMS, NoSQL, ...
   - Spark can get live data (streaming) -- Kafka, Flume, Kinesis, ...
2. Clean & inspect the data.
   - Spark can do Regex, Filtering, Corrections, Enriching, etc.
   - Spark can do Batch processing and/or Streaming processing
3. Feature Engg -- Extract features into numerical vectors
   - Spark has Transformers and Estimators.
4. Build ML model using portion of data (training set).
   - Spark ML supports Supervised ML, Unsupervised ML, Recommendations, NLP, etc.
5. Evaluate model using portion of data (test set).
   - Spark does accuracy check using Evaluators.

SUNBEAM INFOTECH

6. Serve ML model to predict, recommend, ...
    - Allows to export model in various formats like PMML, Pickle, etc.

Spark ML components

- Includes data collection, cleaning, feature engg, training, evaluating large scale supervised & unsupervised models.
- Advantages/Applications
    - Preprocessing & feature engg.
    - Building models for huge training data.
- (High Level) Components
    - Transformers -- transform()
    - Estimators -- fit() + transform()
    - Evaluators -- checking accuracy of model on test data
    - Pipelines -- stages to build model (includes Transformers & Estimators).
- (Low Level) Components
    - Vectors -- Sparse or Dense vectors.

HBase

```
help

version

status

list
```

```
create 'students', 'name', 'marks'

list
```

SUNBEAM INFOTECH

```
describe 'students'
```

```
create 'books', 'name', 'author', 'subject', 'price'

# put 'tablename', 'row-id', 'column-family', 'cell-value'
put 'books', '1', 'name', 'Atlas Shrugged'
put 'books', '1', 'author', 'Ayn Rand'
put 'books', '1', 'subject', 'Novell'
put 'books', '1', 'price', '523.23'

put 'books', '2', 'name', 'The Fountainhead'
put 'books', '2', 'author', 'Ayn Rand'
put 'books', '2', 'subject', 'Novell'
put 'books', '2', 'price', '432.73'

put 'books', '3', 'name', 'The Alchemist'
put 'books', '3', 'author', 'Paulo Cohelo'
put 'books', '3', 'subject', 'Novell'
put 'books', '3', 'price', '721.3'

put 'books', '4', 'name', 'The Archer'
put 'books', '4', 'price', '345.23'

get 'books', '1', 'name'
get 'books', '1', 'author'
get 'books', '1', 'price'

get 'books', '2', 'name', 'author', 'price'

get 'books', '3'

scan 'books'
```

SUNBEAM INFOTECH

```
disable 'books'

drop 'books'
```

```
create 'contacts', 'name', 'phone', 'email'

list

describe 'contacts'

put 'contacts', '001', 'name:fname', 'Nilesh'
put 'contacts', '001', 'name:lname', 'Ghule'
put 'contacts', '001', 'phone:mobile', '9527331338'
put 'contacts', '001', 'email:office', 'nilesh@sunbeaminfo.com'
put 'contacts', '001', 'email:personal', 'nilesh.testemail@gmail.com'

put 'contacts', '007', 'name:fname', 'James'
put 'contacts', '007', 'name:lname', 'Bond'
put 'contacts', '007', 'phone:mobile', '9422012345'

put 'contacts', '002', 'name:fullname', 'Sunbeam Infotech'
put 'contacts', '002', 'phone:fax', '020-24260308'
put 'contacts', '002', 'email:office', 'siit@sunbeaminfo.com'
put 'contacts', '002', 'email:website', 'www.sunbeaminfo.com'

scan 'contacts'

get 'contacts', '007', 'phone:mobile'

put 'contacts', '007', 'phone:mobile', '9822012345'

get 'contacts', '007'

describe 'contacts'
```

SUNBEAM INFOTECH

```
scan 'contacts', { STARTROW=>'001', ENDROW=>'005' }

scan 'contacts', { COLUMNS=>['name'] }

scan 'contacts', { COLUMNS=>['name:fname', 'phone:mobile'] }

scan 'contacts', { STARTROW=>'001', ENDROW=>'005', COLUMNS=>['name:fname', 'phone:mobile'] }

delete 'contacts', '007', 'name:lname'

get 'contacts', '007'

deleteall 'contacts', '007'

get 'contacts', '007'

disable 'contacts'
```

SUNBEAM INFOTECH