Analytical Operations

After this video you will be able to..

- List common analytical operations within big data pipelines.
- Describe sample applications for these analytical operations.

Analytical Operations

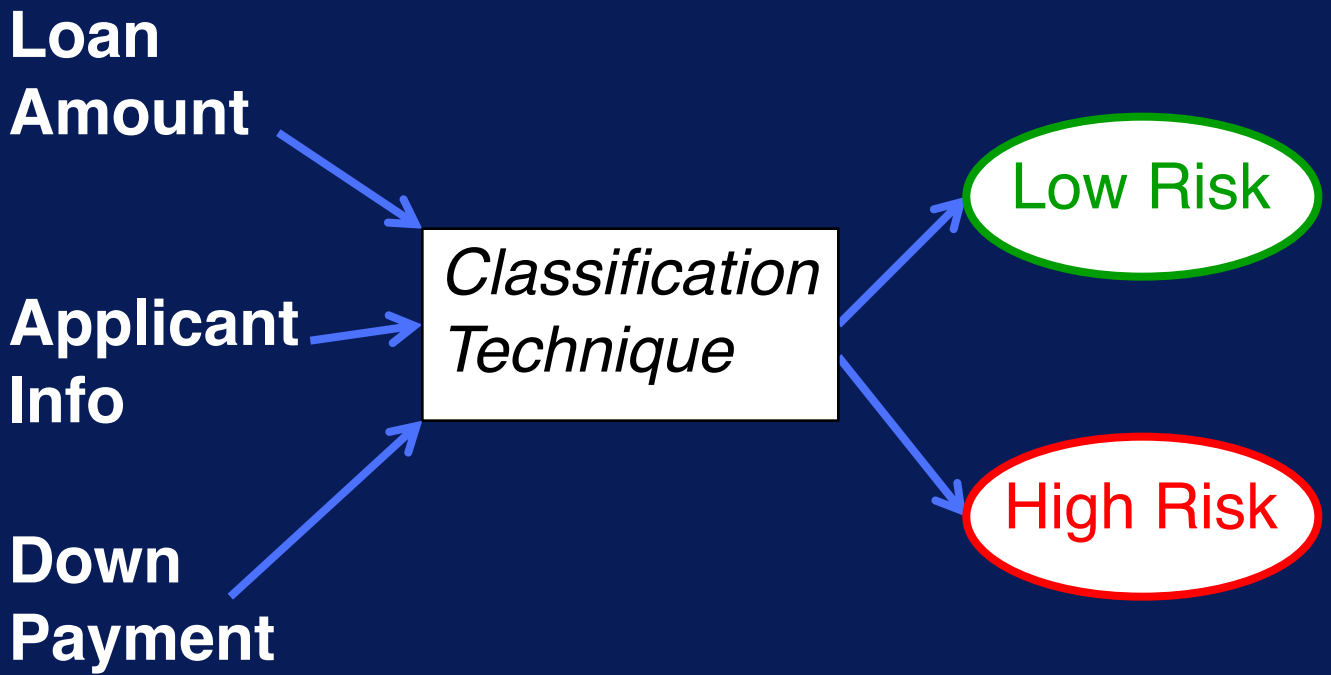
- **Purpose**
  - Discover meaningful trends and patterns in data
  - Gain insights into problem
  - Make data-driven decisions

Sample Analytical Operations

- Classification
- Clustering
- Path analysis
- Connectivity analysis

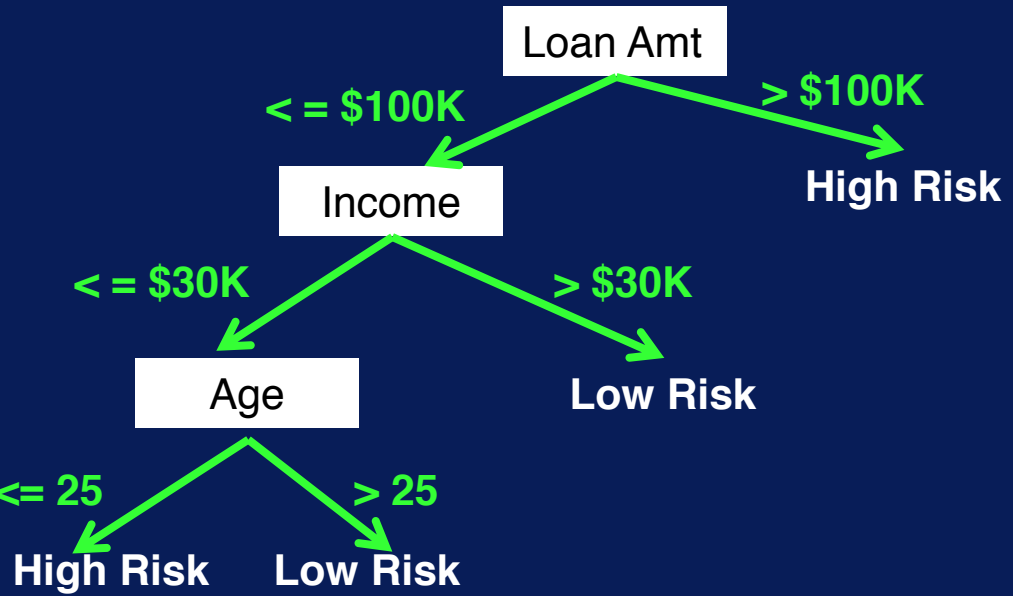
Classification

- **Classify loan application risk**



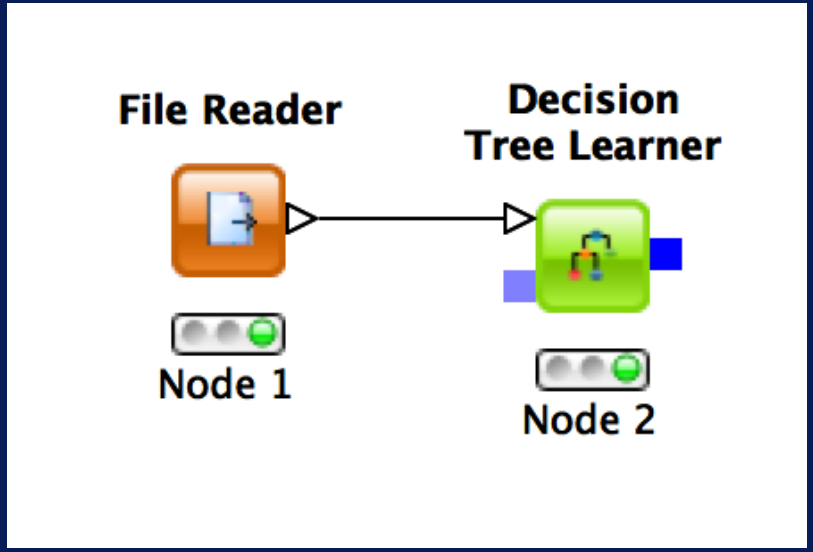
Classification – Decision Tree

- One analytical technique for classification
- Decisions modeled as a tree



Decision Tree in KNIME

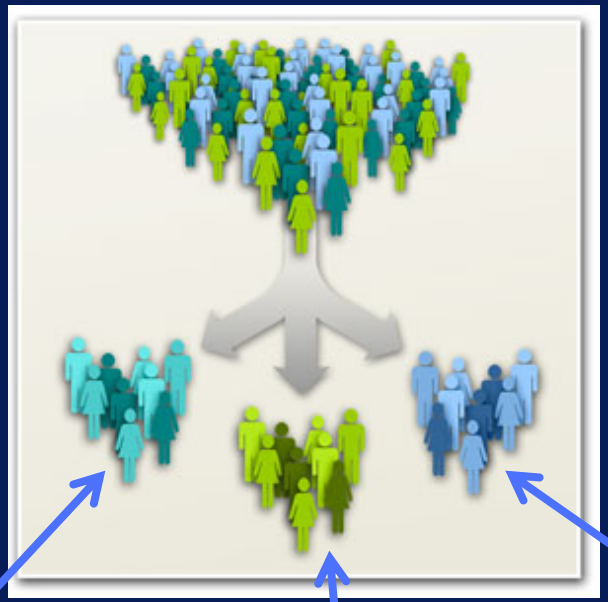
- KNIME workflow for building decision tree from input data



Classification Examples

- Predict whether tumor cells are benign or malignant
- Categorize handwritten digits
- Determine whether credit card transaction is legitimate or fraudulent
- Classify loan application as low-, medium-, or high-risk.

Cluster Analysis



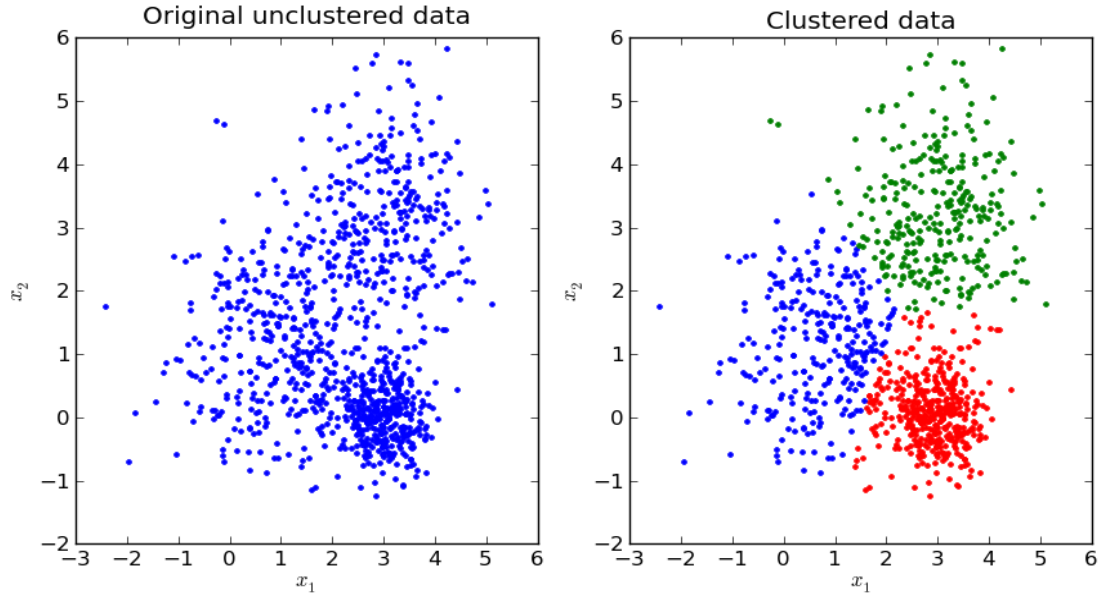
Sci-Fi

Drama

Horror

Cluster Analysis – k-Means

- **K-Means Clustering**
  - Group samples into k clusters



K-Means in Spark

- **Spark Python code for performing k-means on data**

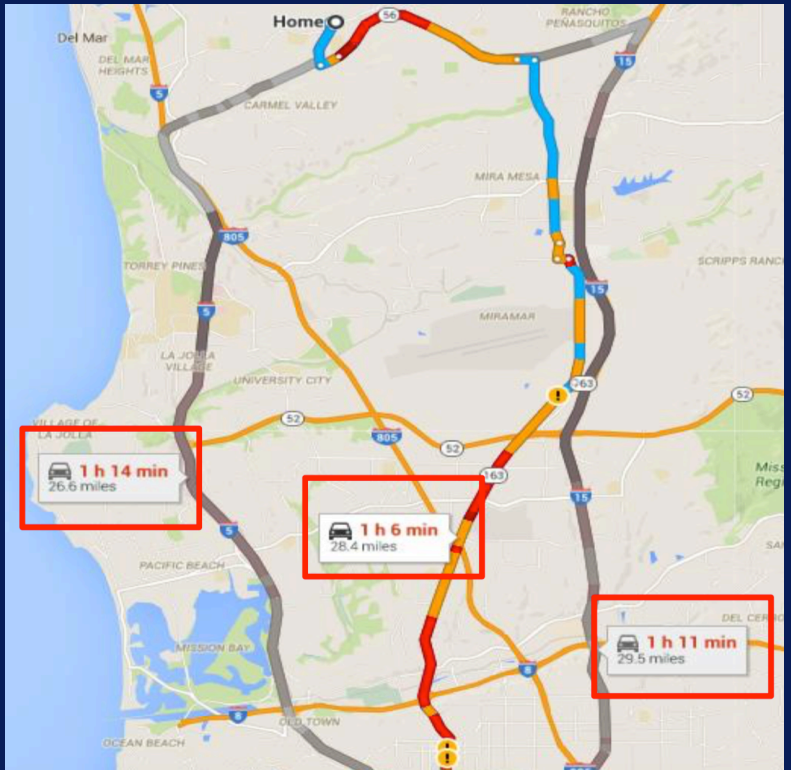
```
# Load and parse the data
data = sc.textFile("data/mllib/kmeans_data.txt")
parsedData = data.map(lambda line:
      array([float(x) for x in line.split(' ')]))

# Cluster the data
clusters = KMeans.train(parsedData, 2,
maxIterations=10, runs=10,
initializationMode="random")
```

Cluster Analysis Examples

- Group customer base into distinct segments
- Find articles or webpages with similar topics
- Identify areas with high incidences of particular crimes
- Determine weather patterns

Path Analysis

Find shortest path from home to work.



Path Analysis

- **Path analysis using Cypher on neo4j**

//Finding shortest path between specific nodes:

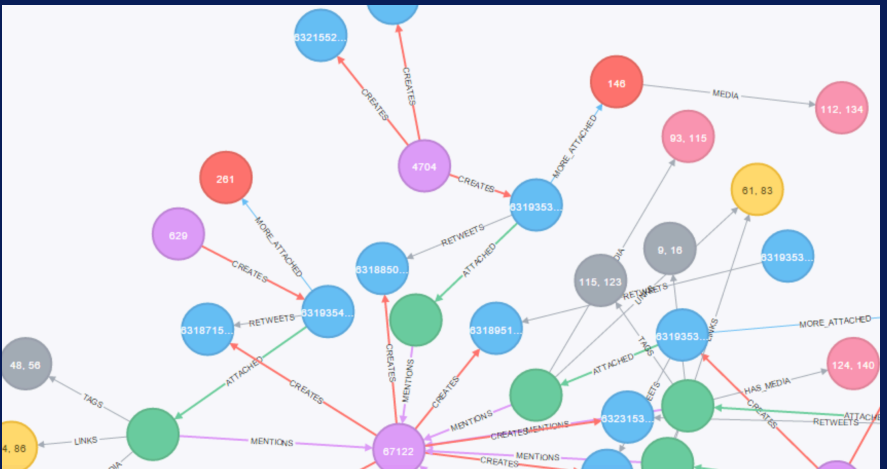
```
match p=shortestPath((a)-[:TO*]-(c))
where a.Name='A' and c.Name='P'
return p, length(p) limit 1
```

//Find all shortest paths:

```
match p = allShortestPaths((source)-[r:TO*]-(destination))
where source.Name='A' and destination.Name = 'P'
return extract(n in nodes (p)| n.Name) as Paths
```

Connectivity Analysis

- **Analyzing tweets**
  - Extract conversation threads
  - Find interacting groups
  - Find influencers in community



Connectivity Analysis

- **Connectivity analysis using Cypher on neo4j**

```
// Find the degree of all nodes
match (n:MyNode)-[r]-()
return n.Name, count(distinct r) as degree
order by degree

// Find degree histogram of the graph
match (n:MyNode)-[r]-()
with n as nodes, count(distinct r) as degree
return degree, count(nodes) order by degree asc
```

Machine Learning Algorithms

- Classification
- Regression
- Cluster Analysis
- Associative Analysis

Graph Analytics Techniques

- Path Analytics
- Connectivity Analytics
- Community Analytics
- Centrality Analytics

Main Take-Aways

- Analytic operations are used to discover meaningful patterns in data to provide insights.
  - e.g.: classification, cluster analysis, path analysis, connectivity analysis

- More analytics are covered in Machine Learning & Graph Analytics courses.