2016-10-25 调研报告

1、目标：测试机器学习算法的稳定性 可靠性 和 扩展性，平台包括spark和flink。

2、测试的方式是：通过数据和配置两个方面来体现

数据：小数据，大数据，稀疏数据，倾斜数据等

配置：partition数量，分片算法，缓冲大小等

3、常用机器学习算法在Spark和flink上实现情况(根据Spark和flink官网文档)

|  |  |  |  |
| --- | --- | --- | --- |
| 常用算法 | | Spark | Flink |
| Classification and regression | **SVMs** | **Y** | **Y** |
| Logistic regression | Y | N |
| Linear regression | Y | N |
| Naive bayes | Y | N |
| Decision tree | Y | N |
| Random forest | Y | N |
| Gradient-boosted trees | Y | N |
| [Multiple linear regression](https://ci.apache.org/projects/flink/flink-docs-release-1.2/dev/libs/ml/multiple_linear_regression.html) | N | Y |
| **[Collaborative filtering](http://spark.apache.org/docs/latest/mllib-collaborative-filtering.html)** | **Alternating least squares(ALS)** | **Y** | **Y** |
| clustering | K-mans | Y | N |
| Gaussian mixture | Y | N |
| Power iteration clustering | Y | N |
| Latent dirichlet allocation | Y | N |
| Bisecting K-means | Y | N |
| Streaming k-means | Y | N |
| KNN | N | Y |
| [Dimensionality reduction](http://spark.apache.org/docs/latest/mllib-dimensionality-reduction.html) | Singular value decomposition(SVD) | Y | N |
| Principal component analysis(PCA) | Y | N |
| Feature extraction and transformation | Feature extraction and transformation | Y | N |
| Frequent pattering mining | [FP-growth](http://spark.apache.org/docs/latest/mllib-frequent-pattern-mining.html" \l "fp-growth) | Y | N |
| [association rules](http://spark.apache.org/docs/latest/mllib-frequent-pattern-mining.html" \l "association-rules) | Y | N |
| [PrefixSpan](http://spark.apache.org/docs/latest/mllib-frequent-pattern-mining.html" \l "prefix-span) | Y | N |