Structural Bioinformatics Training Workshop & Hackathon 2018

mmtfPyspark Advanced

Peter Rose
Director, Structural Bioinformatics Laboratory
San Diego Supercomputer Center
UC San Diego

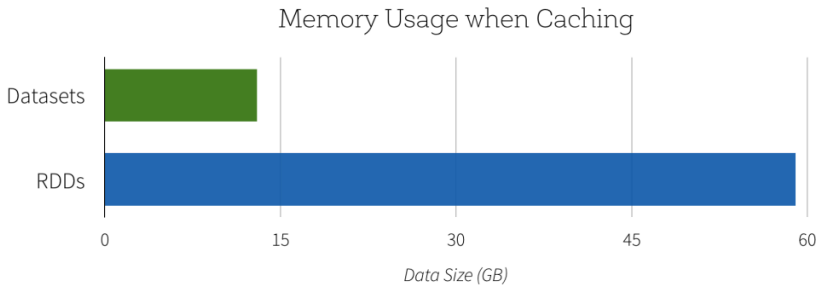
SDSC SAN DIEGO SUPERCOMPUTER CENTER RCSB PDB UC San Diego

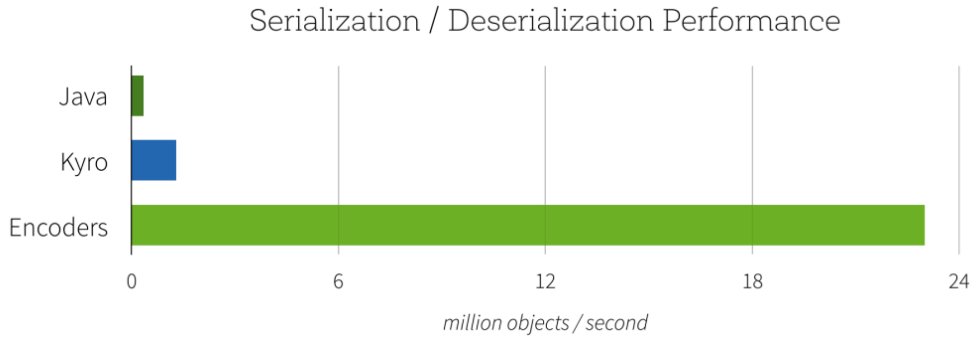
Introduction

- **Spark SQL and Dataset API**
  - Augmenting MMTF data with metadata from PDB and other 3rd party resources
  - Creating datasets of molecular interactions
  - Querying and analyzing datasets

Spark Dataset

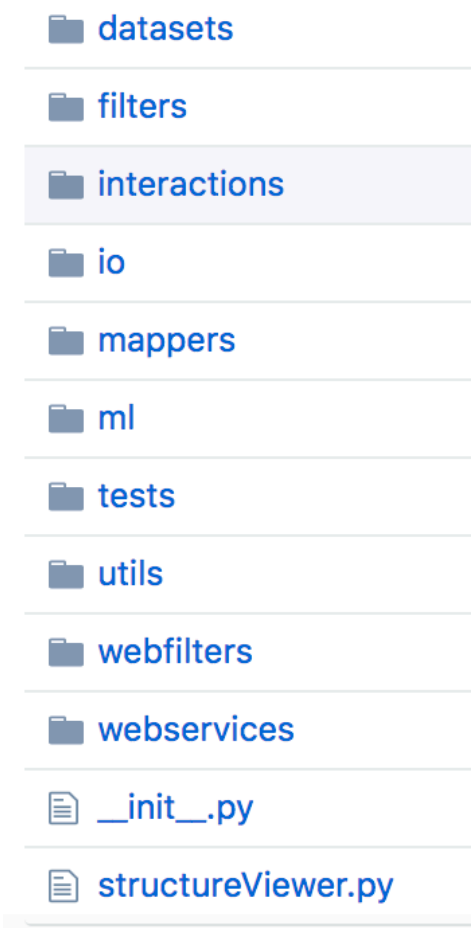
- Table of typed objects with a relational schema
- Similar to Python Pandas and R Dataframes
- Distributed data structure optimized for performance
- Distributed SQL queries on Dataset (Spark SQL)





Source: https://databricks.com/blog/2016/01/04/introducing-apache-spark-datasets.html

mmtfPyspark Modules Covered



- **datasets**
  - Datasets calculated from structure
  - Metadata retrieved from external resources
- **interactions**
  - Ligand-polymer interactions
  - Polymer-polymer interactions

Jupyter Notebook Tutorials

https://github.com/sbl-sdsc/mmtf-workshop-2018/tree/master/4-mmtf-pyspark-advanced

- 1-Metadata
- 2-JoiningDatasets
- 3-MutationsToStructure
- Problem-1 (Solution-1)
- 4-CreateDatasets
- Problem-2 (Solution-2)

Summary

- **Spark Dataset API provides an efficient distributed tabular data structure**
- **Can be queried using Spark SQL**
- **We used datasets to**
  - get additional metadata not available in MMTF
  - store and query the results of structural calculations

Resources

- **Spark SQL, DataFrames and Datasets Guide**
  - https://spark.apache.org/docs/latest/sql-programming-guide.html
- **MMTF Website**
  - https://mmtf.rcsb.org
- **GitHub Repository**
  - https://github.com/sbl-sdsc/mmtf-pyspark
  - https://github.com/sbl-sdsc/mmtf-spark

- **RCSB PDB Web Services and Query System**
  - Rose, PW, et al. (2013) The RCSB Protein Data Bank: new resources for research and education, Nucleic Acids Res 41: D475-D482. https://doi.org/10.1093/nar/gks1200
  - Rose, PW, et al. (2011) The RCSB Protein Data Bank: redesigned web site and web services, Nucleic Acids Res 39: D392-D401. https://doi.org/10.1093/nar/gkq1021

Funding

This workshop was supported by the National Cancer Institute of the National Institutes of Health under Award Number U01CA198942. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.