



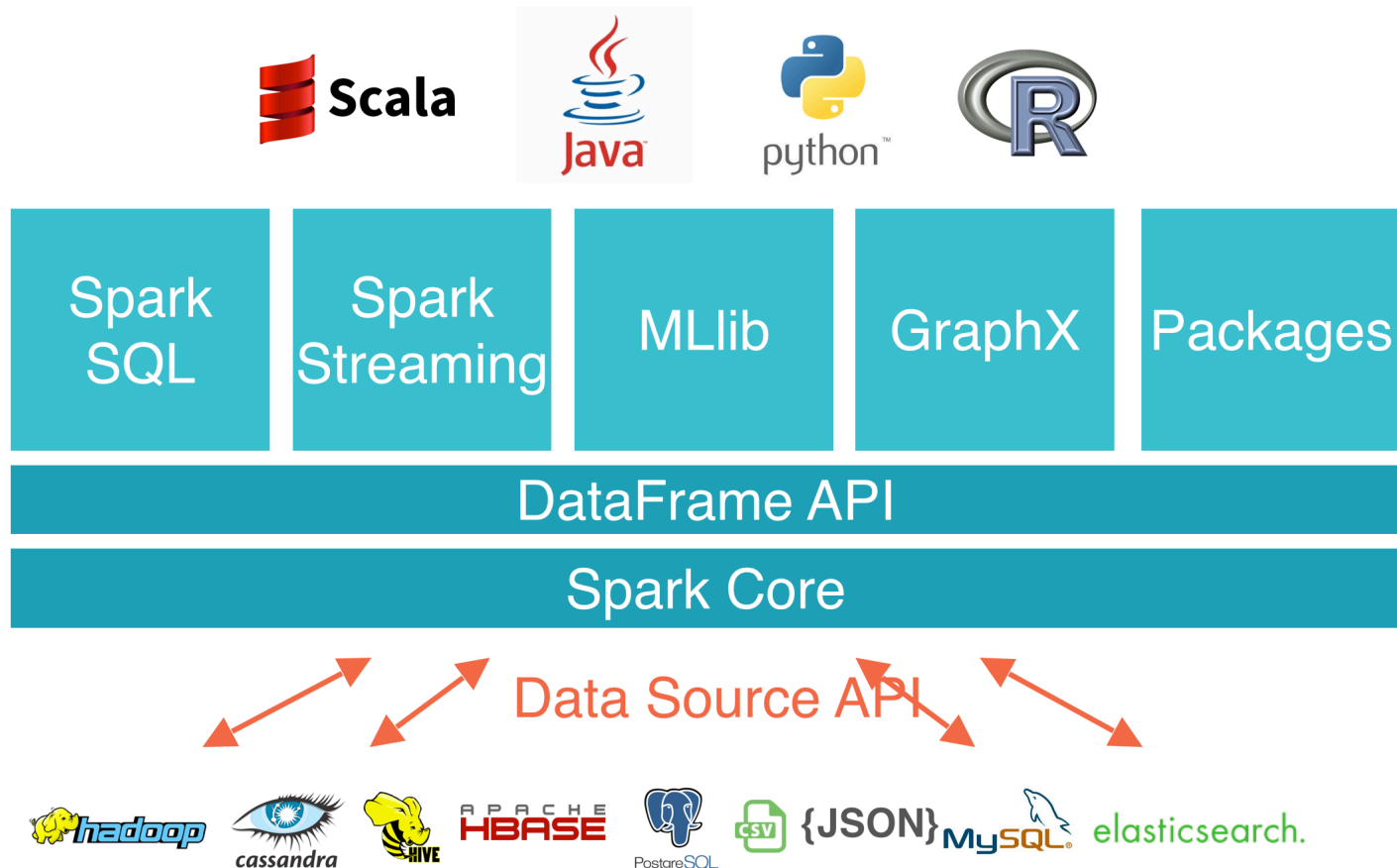
Structural Bioinformatics Training Workshop & Hackathon 2018

mmtfPyspark

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

Spark Ecosystem

Apache Spark is a fast and general engine
for large-scale data processing



What is mmtfPyspark?

- A framework for interactive analysis and mining of 3D macromolecular structures
- Powered by MMTF (MacroMolecular Transmission Format), a compact data format that facilitates efficient network transfer and high-performance parsing and processing of 3D structures
- Built on Apache Spark, a framework for distributed, parallel in-memory processing
- Uses Spark-SQL for queries and Spark-Mllib for machine learning
- Available with Java and Python API

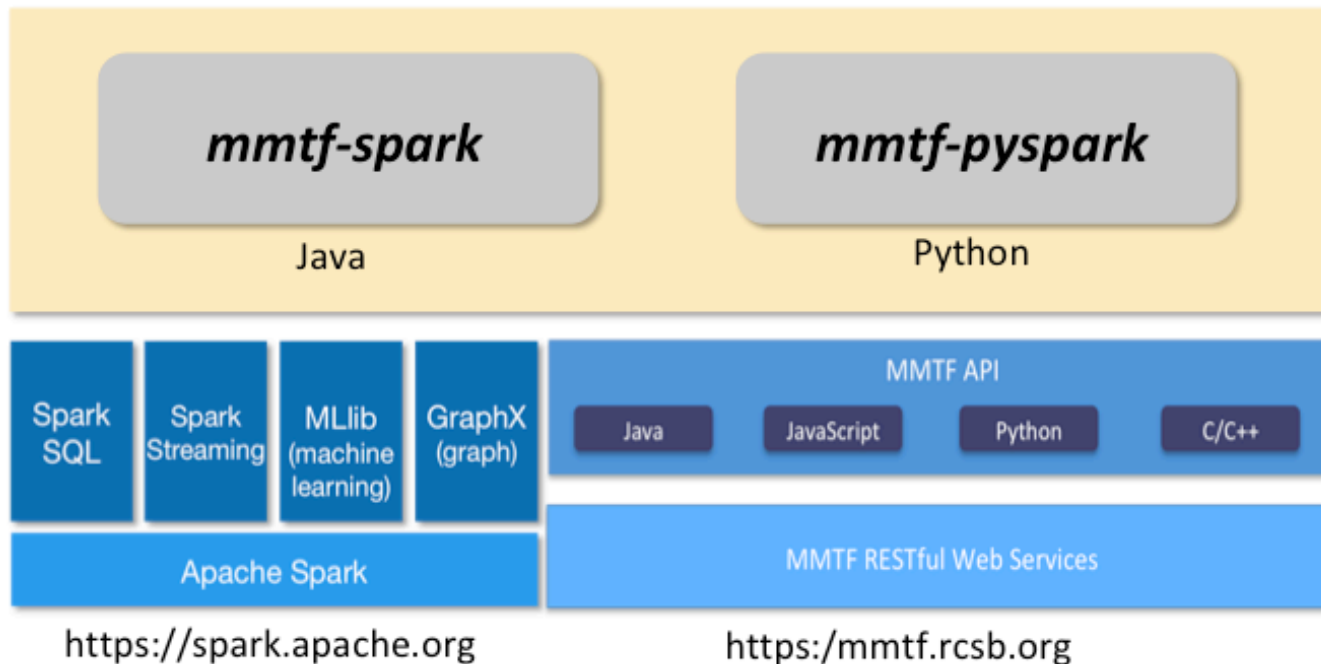
MMTF-Spark

mmtf-spark (Java)

- High performance processing
- Suitable for large-scale calculations
- Integration with other libraries, e.g., BioJava

mmtf-pyspark (Python)

- Interactive scripting
- 2D and 3D visualization
- ML/DL tool ecosystem
- Sharable data analysis in Jupyter Notebooks



mmtfPyspark Modules Covered

datasets

filters

interactions

io

mappers

ml

tests

utils

webfilters

webservice

__init__.py

structureViewer.py

- io
 - Read MMTF structures into Spark RDDs
 - Write MMTF structures
- filters
 - Filter RDDs by structural characteristics and metadata
- webfilters
 - Filter using metadata from external resources
- mappers
 - Map structures to different representations

MMTF Data Sources

- **Download single MMTF files using web services**
 - Analyze 10s – 100 PDB entries
 - <https://mmtf.rcsb.org/v1.0/full/4HHB.mmtf.gz>
- **Download MMTF Hadoop Sequence files**
 - Analyze 1000s or all PDB entries
 - <https://mmtf.rcsb.org/v1.0/hadoopfiles/full.tar>
 - <https://mmtf.rcsb.org/v1.0/hadoopfiles/reduced.tar>
- **Info about downloading**
 - <https://mmtf.rcsb.org/download.html>







Hadoop “Sequence” Files

- **A flat file of binary key/value pairs**
- **Used by Big Data Frameworks (Hadoop, Spark)**
 - File systems need few big files for efficient processing
- **Files are splittable**
 - Can be processed in parallel
- **Often consists of a directory of Sequence files**
- **See <https://wiki.apache.org/hadoop/SequenceFile>**

MMTF Hadoop Sequence Files

- **Two representations**
 - **full**
 - all atoms
 - full data precision
 - **reduced**
 - polymers
 - polypeptides: C-alpha
 - polynucleotides: P
 - 1st model only (e.g., NMR)
 - no alternative locations
 - except polysaccharides
 - » all atom
 - non-polymers
 - all atoms
 - water
 - excluded
 - Reduced precision (0.1):
coordinates, temperature-factor,
occupancy

- **Example: full directory structure**

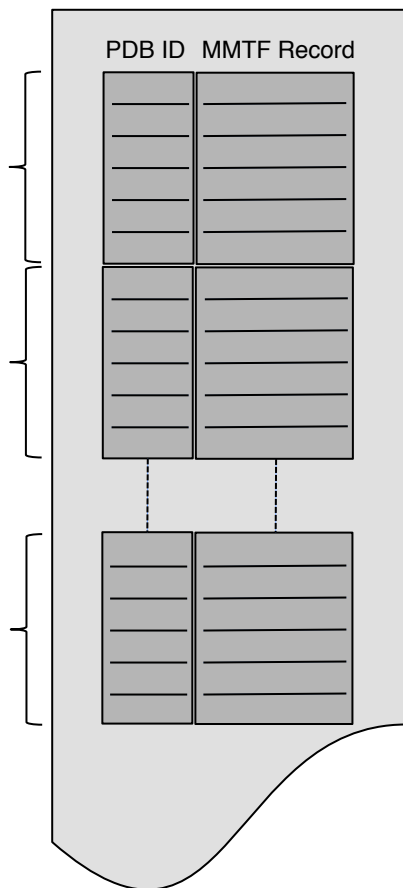
Name	^	Date Modified	Size
 _2017-06-06.txt		Jun 6, 2017, 5:02 PM	Zero bytes
 _SUCCESS		Jun 2, 2017, 2:07 PM	Zero bytes
 part-00000		Jun 2, 2017, 2:00 PM	9.8 MB
 part-00001		Jun 2, 2017, 2:00 PM	13.9 MB
 part-00002		Jun 2, 2017, 2:00 PM	33.3 MB
 part-00003		Jun 2, 2017, 2:00 PM	33.4 MB

- **Timestamp file (release date)**
 - __yyyy-mm-dd.txt
- **Updated every Wed. ~00:00 UTC**
- **Multiple sequence files**
 - part-00000 ...
- **Download**
 - <https://mmtf.rcsb.org/download.html>

MMTF-Spark Data Pipeline

MMTF Hadoop Sequence File
(directory in Spark)

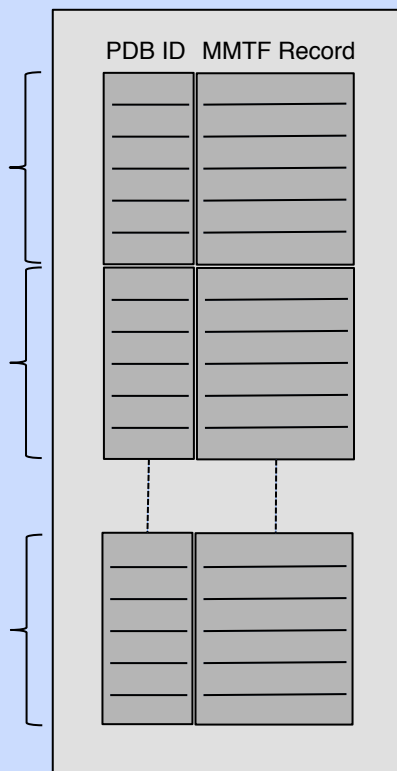
Splittable
Hadoop
Sequence
file enables
parallel I/O



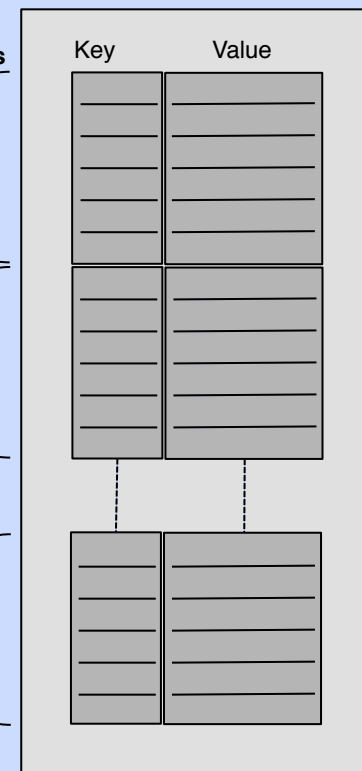
Parallel I/O
(e.g., using
HDFS)

Partitions
distributed
over
multiple
cores and
servers

SPARK RDD
(Resilient Distributed Dataset)



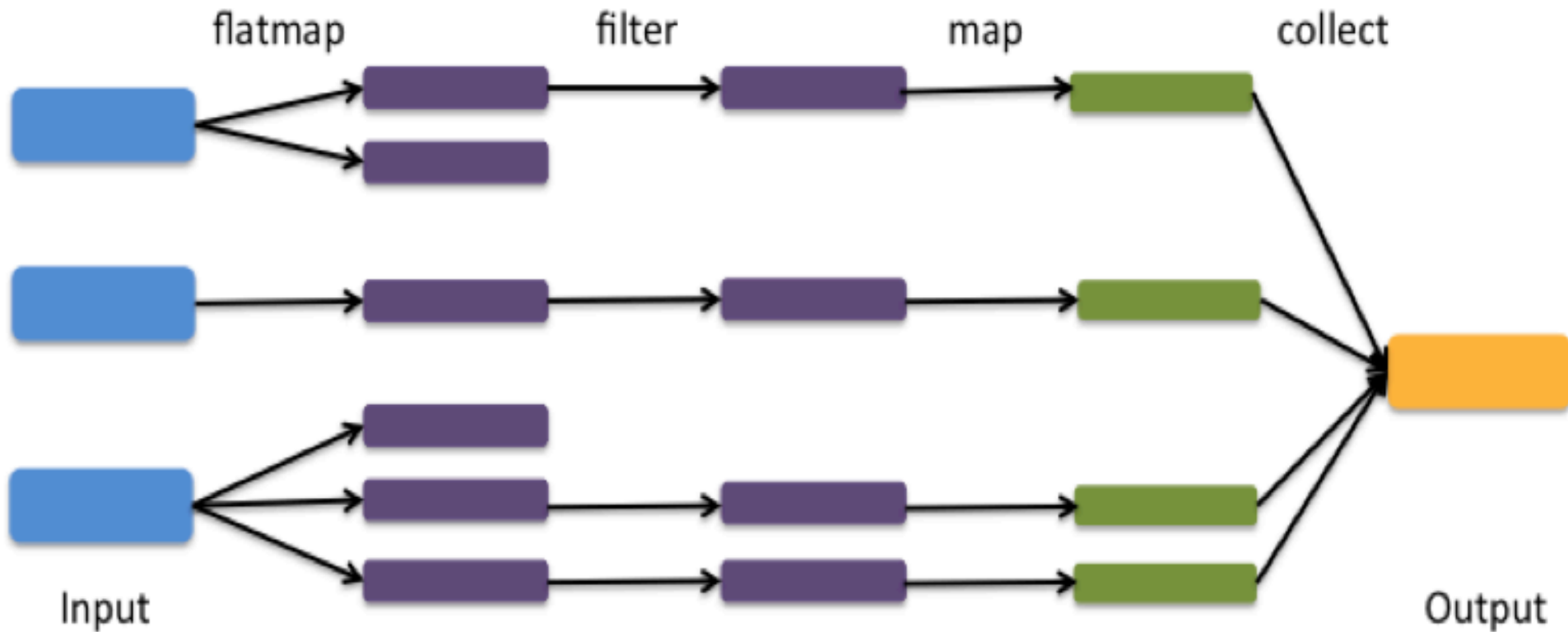
**Parallel
Transformations**



MMTF Hadoop Sequence Files for Workshop

- Sample MMTF Hadoop Sequence Files are included in the workshop repository
 - <https://github.com/sbl-sdsc/mmtf-workshop-2018/tree/master/resources>
- The sample files contain 9756 representative PDB structures
 - mmtf_full_sample
 - mmtf_reduced_sample

Example of a Spark Workflow



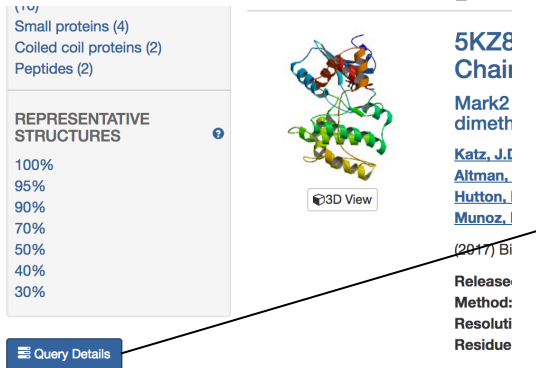
Jupyter Notebook Tutorials

<https://github.com/sbl-sdsc/mmtf-workshop-2018/tree/master/3-mmtf-pyspark>

- 1-Input
- 2-Filtering
- **Problem-1 (Solution-1)**
- 3-Webfiltering
- 4-Flatmapping
- 5-MapReduce
- **Problem-1 (Solution-2)**
- 6-Output
- 7-OutputTo3DViewer

Filtering Using AdvancedQuery

- Run any advanced query at <http://www.rcsb.org>
- Go to results page and click Query Details



Query in XML format

```
<orgPdbQuery>
  <version>head</version>
  <queryType>org.pdb.query.simple.StoichiometryQuery</queryType>
  <description>Stoichiometry in biological assembly: Stoichiometry is A3B3C3</description>
  <queryId>FFD43033</queryId>
  <resultCount>96</resultCount>
  <runtimeStart>2018-05-06T23:20:25Z</runtimeStart>
  <runtimeMilliseconds>280</runtimeMilliseconds>
  <stoichiometry>A3B3C3</stoichiometry>
</orgPdbQuery>
```

```
query = (
    "<orgPdbQuery>"
    "<queryType>org.pdb.query.simple.StoichiometryQuery</queryType>"
    "<stoichiometry>A3B3C3</stoichiometry>"
    "</orgPdbQuery>"
)
trimer_of trimers = pdb.filter(AdvancedQuery(query))
```

Filtering by SMILES

```
// keep structures that contain a chemical component  
// with this substructure  
pdb = pdb.filter(  
    ChemicalStructureQuery("OC(=O)CCCC[C@@H]1SC[C@@H]2N  
    C(=O)N[C@H]12",  
    ChemicalStructureQuery.SUBSTRUCTURE, 0))
```

```
// keep structures that contain a chemical component  
// that is >= 70% similar to the query structure  
similarity = 70  
pdb = pdb.filter(  
    ChemicalStructureQuery("OC(=O)CCCC[C@@H]1SC[C@@H]2N  
    C(=O)N[C@H]12",  
    ChemicalStructureQuery.SIMILAR, similarity))
```

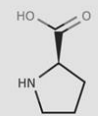
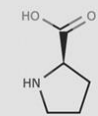
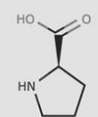
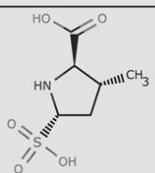
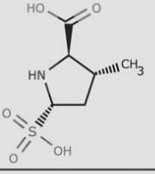
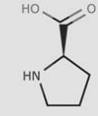
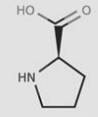
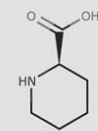
Query Types:

EXACT

SUBSTRUCTURE

SUPERSTRUCTURE

SIMILAR

Search type	Query	Result
Exact		
Substructure		
Superstructure		
Similar		

Summary

- **mmtfPyspark: Framework for parallel distributed mining of the PDB with Apache Spark**
- **MMTF Hadoop Sequence file is an efficient container format to process large number of structures**
- **PDB structures represented as key/value pairs**
- **Spark transformations**
 - filter, keys, map, flatMap
- **Spark actions**
 - count, reduce, collect

Resources

- **MMTF Website**

- <https://mmtf.rcsb.org>

- **Git Repositories**

- <https://github.com/sbl-sdsc/mmtf-pyspark>
- <https://github.com/sbl-sdsc/mmtf-spark>
- <https://github.com/sbl-sdsc/mmtf-workshop-2017>

- **MMTF File Format**

- Bradley AR, et al. (2017) MMTF—An efficient file format for the transmission, visualization, and analysis of macromolecular structures. PLOS Computational Biology 13(6): e1005575.
<https://doi.org/10.1371/journal.pcbi.1005575>
- Valasatava Y, et al. (2017) Towards an efficient compression of 3D coordinates of macromolecular structures. PLOS ONE 12(3): e0174846.
<https://doi.org/10.1371/journal.pone.0174846>

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.

