Exploring Tachyon and Spark

Kenny Gorman
Chief Technologist for Data at Rackspace
Co-Founder, ObjectRocket

@rackspace @kennygorman

Rackspace Tech Research

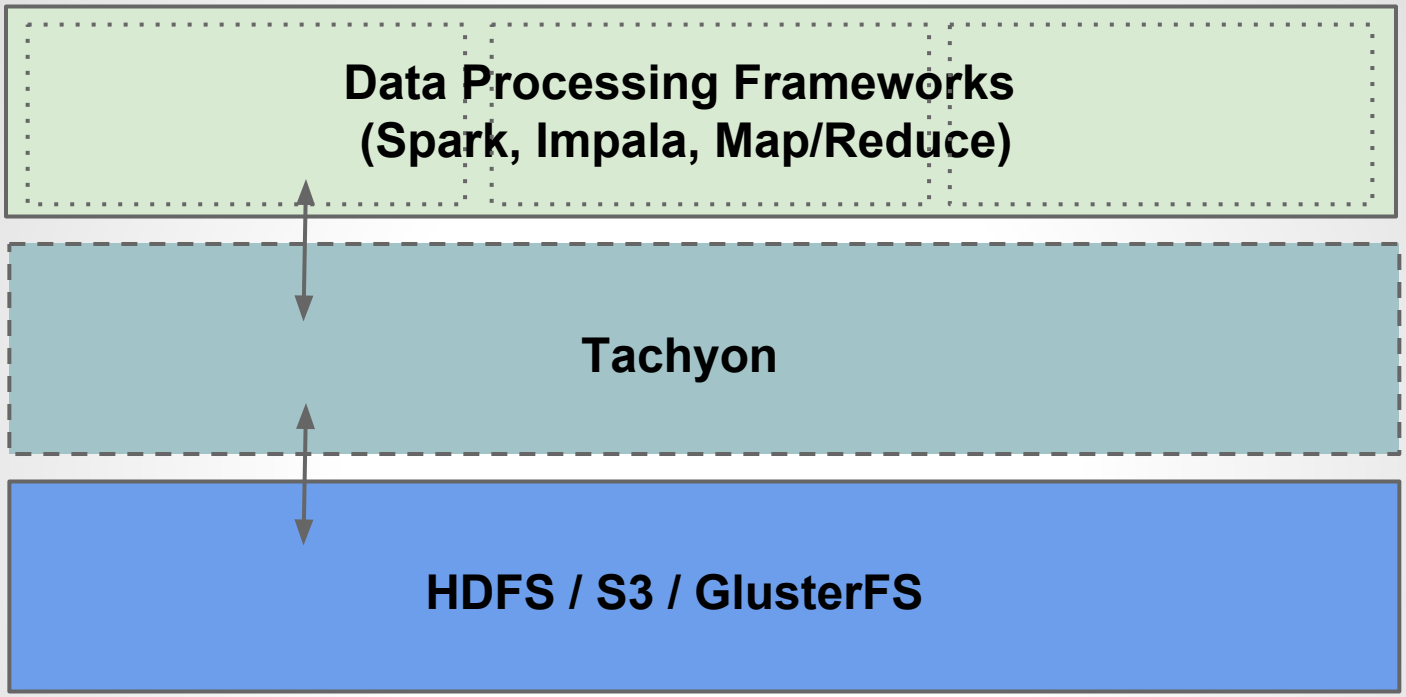
- Constantly looking for new technologies
- Improve our offerings
- Get our hands on the newest stuff, generate opinions
- Some become products or services
- Enter Tachyon

What is Tachyon?

"Storage level component providing memory speed storage access for data processing frameworks"

- Developed at UC Berkeley AMPLab
- Apache 2.0 License
- Memory centric file system front end
- Implements HDFS, so usable in HDFS based systems
- v0.6.0 https://github.com/amplab/tachyon
- Command line interface, and Web UI

Basic Tachyon Architecture



- Masters and workers
- Zookeeper

Benefits

"Reliable data sharing at memory-speed within and across cluster frameworks/jobs"

- Access at memory speed
- Keep one copy of data in memory, each spark task has access
- On crash, copy of data is lost in block manager
- Less GC

Using Tachyon

Simple I/O with HDFS:

```
// load some stuff

$ val s = sc.textFile("tachyon-ft://stanbyHost:19998/mydataset.txt")
$ val x = s.count()

// save some output
$ x.saveAsTextFile("tachyon-ft://activeHost:19998/mydataset_tachyon.txt")
```

With Spark RDD:

```
// load some data

$ val rdd = sc.textFile(“htfs://users/interesting_stuff.json”)

// make an RDD
$ rdd.persist(StorageLevel.OFF_HEAP)
```

Tachyon + Rackspace

- Use a CBD Spark cluster at https://mycloud.rackspace.com
- Spark OnMetal
- Compile from SNAPSHOT. It's early. ;-).
- Simple master on gateway node.
- Install maven and jdk
- Compile for Hortonworks HDP 2.1 Hadoop
- Configuration file changes
- Startup/Format
- Run tests! Enjoy!

HOWTO:

https://gist.github.com/kgorman/f23a509183aecb02bfac

Important links:

http://tachyon-project.org/master/
http://www.cs.berkeley.edu/~haoyuan/talks/Tachyon_2014-10-16-Strata.pdf

Contact

@kennygorman

@rackspace

kenny.gorman@rackspace.com

http://developer.rackspace.com/databases/