The other Apache Technologies your Big Data solution needs!

Nick Burch
CTO, Quanticate

Quanticate
A PASSION FOR EXCELLENCE

DATA MANAGEMENT • BIOSTATISTICS • PROGRAMMING • MEDICAL WRITING • PHARMACOVIGILANCE

Quanticate

A PASSION FOR EXCELLENCE

Nick Burch
CTO, Quanticate

Global Solutions from the World's Largest Data-Dedicated CRO

www.quanticate.com

The Apache Software Foundation

Apache Technologies as in the ASF

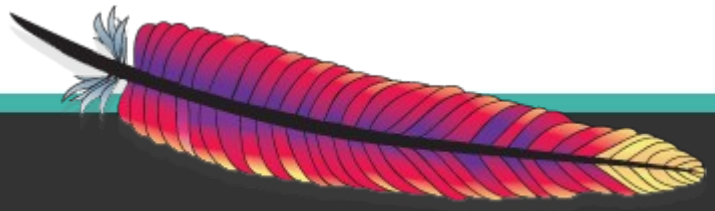
154 Top Level Projects

33 Incubating Projects (> 100 past ones)

Y is the only letter we lack

A, C, O, S & T are popular letters, >= 12 each

Meritocratic, Community driven
Open Source



Quanticate
A PASSION FOR EXCELLENCE

A Growing Foundation...

November 2014:

- 154 Top Level Projects
- 33 Incubating Projects

June 2013:

- 117 Top Level Projects
- 37 Incubator Projects

June 2011:

- 91 Top Level Projects
- 59 Incubating Projects

Quanticate

A PASSION FOR EXCELLENCE

What we're not covering

Quanticate

A PASSION FOR EXCELLENCE

Projects not being covered include

Cassandra	cassandra.apache.org
CouchDB	couchdb.apache.org
Flume	flume.apache.org
Giraph	giraph.apache.org
Hadoop	hadoop.apache.org
Hive	hive.apache.org
HBase	hbase.apache.org

Quanticate
A PASSION FOR EXCELLENCE

Projects not being covered include

Lucene	lucene.apache.org
Mahout	mahout.apache.org
Mesos	mesos.apache.org
Nutch	nutch.apache.org
OODT	oodt.apache.org
Spark	
Storm	etc!

Quanticate

A PASSION FOR EXCELLENCE

What we are looking at

Quanticate

A PASSION FOR EXCELLENCE

Talk Structure

New Things from the Incubator

Established Big Data projects not to forget about

Non Big-Data projects that can help flesh out an overall solution

Many projects to cover – this is only an overview!

Quanticate

A PASSION FOR EXCELLENCE

Audience Participation:
A show of hands...

Quanticate

A PASSION FOR EXCELLENCE

Making the most of the talk

- Slides are available on the Conference site
- Lots of information and projects to cover
- Just an overview, see project sites for more
- Try to take note / remember the projects you think will matter to you
- Don't try to take notes on everything – there's too much!

Quanticate

A PASSION FOR EXCELLENCE

What's new(ish) in the Apache Incubator?

Quanticate

A PASSION FOR EXCELLENCE

Argus : argus.incubator.apache.org

- Data Security framework for Hadoop
- Central administration of all security related tasks, with central UI + Rest API
- Fine-grained model controlling access to data, components, actions, operations
- Centralised Auditing and Monitoring
- Role based, attribute based etc
- Standardised authz for Hadoop

Quanticate
A PASSION FOR EXCELLENCE

Aurora : aurora.incubator.apache.org

- Service Scheduler on top of Mesos
- To run+manage long-running services
- Deployment and scheduling of jobs
- Uses a DSL to define services
- Heath checking, failure monitoring, restarting etc
- Aurora jobs are made up of many different Mesos tasks

Quanticate
A PASSION FOR EXCELLENCE

Blur : incubator.apache.org/blur

Search engine for massive amounts of structured data at high speed

Query rich, structured data model

US Census example: show me all of the people in the US who were born in Alaska between 1940 and 1970 who are now living in Kansas.

Built on Apache Hadoop

Quanticate

A PASSION FOR EXCELLENCE

Brooklyn : brooklyn.incubator.a.o

- Framework for modelling, monitoring and managing applications from blueprints
- Blueprints describe app from components, many built in, using bash, Java, Chef etc
- Deployed across multiple machines automatically: cloud, private, docker etc
- Scale, replace, restart etc
- Metrics monitored

Quanticate
A PASSION FOR EXCELLENCE

Calcite : calcite.incubator.a.o

- Dynamic Data Management framework
- Highly customisable engine for planning and parsing queries on data from a wide variety of formats
- SQL interface for data not in relational databases, with query optimisation
- Complementary to Hadoop and NoSQL systems, esp. combinations of them
- Formerly known as Optiq

Quanticate
A PASSION FOR EXCELLENCE

DataFu : datafu.incubator.apache.org

- Collection of libraries for working with large-scale data in Hadoop, for data mining, statistics etc
- Provides Map-Reduce jobs and high level language functions for data analysis, eg statistics calculations
- Incremental processing with Hadoop with sliding data, eg computing daily and weekly statistics

Quanticate
A PASSION FOR EXCELLENCE

Drill : incubator.apache.org/drill

- Drill is a distributed system for interactive analysis of large-scale datasets, inspired by Google's Dremel
- Low-latency queries natively on rapidly evening multi-structured datasets
- Aiming to be able to scale to 10k+ servers, petabytes or data and trillions of records, all within seconds

Quanticate
A PASSION FOR EXCELLENCE

Falcon : falcon.incubator.apache.org

- Data management and processing framework built on Hadoop
- Quickly onboard data + its processing into a Hadoop based system
- Declarative definition of data endpoints and processing rules, inc dependencies
- Orchestrates data pipelines, management, lifecycle, motion etc

Quanticate
A PASSION FOR EXCELLENCE

Flink : flink.incubator.apache.org

Flink is an open source system for expressive, declarative, fast, and efficient data analysis. Flink combines the scalability and programming flexibility of distributed MapReduce-like platforms with the efficiency, out-of-core execution, and query optimization capabilities found in parallel databases.

Quanticate
A PASSION FOR EXCELLENCE

Ignite : ignite.incubator.apache.org

- Formerly known as GainGrid
- Only just entered incubation
- In-Memory data fabric
- High performance, distributed data management between heterogeneous data sources and user applications
- Stream processing and compute grid
- Structured and unstructured data

Quanticate
A PASSION FOR EXCELLENCE

MetaModel: metamodel.incubator.a.o

MetaModel is a data access framework, providing a common interface for exploration and querying of different types of datastores

Safe & Uniform model for querying datastores

Uses native support where available

Implements it on top for other datastores

RDBMS, CSV, NoSQL, XML, XLS, JSON etc

Quanticate

A PASSION FOR EXCELLENCE

MRQL : mrql.incubator.apache.org

- Large scale, distributed data analysis system, built on Hadoop, Hama, Spark
- Query processing and optimisation
- SQL-like query for data analysis
- Works on raw data in-situ, such as XML, JSON, binary files, CSV
- Powerful query constructs avoid the need to write MapReduce code
- Write data analysis tasks as SQL-like

Quanticate
A PASSION FOR EXCELLENCE

Parquet : parquet.incubator.a.o

- Columnar storage format for Hadoop
- Compressed, efficient columnar data representation for any Hadoop projects
- Allows complex nested data structures
- Based on record shredding/assembly algorithm from the Dremel Paper
- Per-column compressions / encodings

Quanticate
A PASSION FOR EXCELLENCE

REEF : reef.incubator.apache.org

REEF (Retainable Evaluator Execution Framework) is a scale-out computing fabric that eases the development of Big Data applications on top of resource managers such as Apache YARN and Mesos.

Quanticate

A PASSION FOR EXCELLENCE

Samza : samza.incubator.apache.org

Samza provides a system for processing stream data from publish-subscribe systems such as Apache Kafka. The developer writes a stream processing task, and executes it as a Samza job. Samza then routes messages between stream processing tasks and the publish-subscribe systems that the messages are addressed to.

Quanticate
A PASSION FOR EXCELLENCE

Sentry : sentry.incubator.apache.org

Sentry is a highly modular system for providing fine grained role based authorization to both data and metadata stored on an Apache Hadoop cluster.

Quanticate

A PASSION FOR EXCELLENCE

Slider : slider.incubator.apache.org

Slider is a collection of tools and technologies to package, deploy, and manage long running applications on Apache Hadoop YARN clusters.

Quanticate

A PASSION FOR EXCELLENCE

Twill : twill.incubator.apache.org

Twill is an abstraction over Apache Hadoop YARN that reduces the complexity of developing distributed applications, allowing developers to focus more on their business logic

Quanticate

A PASSION FOR EXCELLENCE

UserGrid : usergrid.incubator.a.o

- Backend-as-a-Service “Baas” “mBaaS”
- Distributed NoSQL database + asset storage
- Mobile and server-side SDKs
- Rapidly build mobile and/or web applications, inc content driven ones
- Provides key services, eg users, queues, storage, queries etc

Quanticate

A PASSION FOR EXCELLENCE

Loading and Querying

Quanticate

A PASSION FOR EXCELLENCE

Pig – pig.apache.org

- Originally from Yahoo, entered the Incubator in 2007, graduated 2008, very widely used
- Provides an easy way to query data, which is compiled into Hadoop M/R (not SQL-like)
- Typically 1/20th of the lines of code, and 1/15th of the development time
- Optimising compiler – often only slower, occasionally faster!

Quanticate
A PASSION FOR EXCELLENCE

Gora – gora.apache.org

- ORM Framework for (NoSQL) Column Stores
- Grew out of the Nutch project
- Supports HBase, Cassanda, Hypertable
- K/V: Voldermort, Redis
- HDFS Flat Files, plus basic SQL
- Data is stored in Avro (more later)
- Query with Pig, Lucene, Hive, Cascading, Hadoop M/R, or native Store code

Quanticate
A PASSION FOR EXCELLENCE

Sqoop – sqoop.apache.org (no u!)

- Bulk data transfer tool for big data systems
- Hadoop (HDFS), HBase and Hive on one side
- SQL Databases on the other
- Can be used to import existing data into your big data cluster
- Or, export the results of a big data job out to your data wharehouse
- Generates code to work with data

Quanticate
A PASSION FOR EXCELLENCE

Building MapReduce Jobs

Quanticate

A PASSION FOR EXCELLENCE

Avro – avro.apache.org

- Language neutral data serialization
- Rich data structures (JSON based)
- Compact and fast binary data format
- Code generation optional for dynamic languages
- Supports RPC
- Data includes schema details

Quanticate

A PASSION FOR EXCELLENCE

Avro – avro.apache.org

Schema is always present – allows dynamic typing and smaller sizes

Java, C, C++, C#, Python, Perl, Ruby, PHP

Different languages can transparently talk to each other, and make RPC calls to each other

Often faster than Thrift and ProtoBuf

No streaming support though

Quanticate

A PASSION FOR EXCELLENCE

Thrift – thrift.apache.org

- Java, C++, Python, PHP, Ruby, Erlang, Perl, Haskell, C#, JS, Cocoa, OCaml and more!
- From Facebook, at Apache since 2008
- Rich data structure, compiled down into suitable code
- RPC support too
- Streaming is available
- Worth reading the White Paper!

Quanticate
A PASSION FOR EXCELLENCE

MRUnit – mrunit.apache.org

- Built on top of JUnit
- Checks Map, Reduce, then combined
- Provides test drivers for Hadoop
- Avoids you needing lots of boiler plate code to start/stop Hadoop
- Avoids brittle mock objects
- Handles multiple input K/Vs
- Counter checking

Quanticate
A PASSION FOR EXCELLENCE

For the Cloud

Quanticate

A PASSION FOR EXCELLENCE

Provider Independent Cloud APIs

- Lets you provision, manage and query Cloud services, without vendor lock-in
- Translates general calls to the specific (often proprietary) ones for a given cloud provider
- Work with remote and local cloud providers (almost) transparently

Quanticate

A PASSION FOR EXCELLENCE

Provider Independent Cloud APIs

- Create, stop, start, reboot and destroy instances
- Control what's run on new instances
- List active instances
- Fetch available and active profiles
- EC2, Eycalyptos, Rackspace, RHEV, vSphere, Linode, OpenStack

Quanticate

A PASSION FOR EXCELLENCE

LibCloud – libcloud.apache.org

- Python library (limited Java support)
- Very wide range of providers
- Script your cloud services

```
from libcloud.compute.types import Provider
from libcloud.compute.providers import get_driver

EC2_ACCESS_ID = 'your access id'
EC2_SECRET_KEY = 'your secret key'

Driver = get_driver(Provider.EC2)
conn = Driver(EC2_ACCESS_ID, EC2_SECRET_KEY)

nodes = conn.list_nodes()
# [<Node: uuid=..., state=3, public_ip=['1.1.1.1'],
    provider=EC2 ...>, ...]
```

Quanticate
A PASSION FOR EXCELLENCE

DeltaCloud – deltacloud.apache.org

- REST API (xml) + web portal
- Major Cloud Providers, RHEV-M, vSphere

```
<instances>
  <instance href="http://fancycloudprovider.com/api/instances/inst1" id='inst1'>
    <owner_id>larry</owner_id>
    <name>Production JBoss Instance</name>
    <image href="http://fancycloudprovider.com/api/images/img3"/>
    <hardware_profile href="http://fancycloudprovider.com/api/hardware_profiles/m1-small"/>
    <realm href="http://fancycloudprovider.com/api/realms/us"/>
    <state>RUNNING</state>
    <actions>
      <link rel="reboot" href="http://fancycloudprovider.com/api/instances/inst1/reboot"/>
      <link rel="stop" href="http://fancycloudprovider.com/api/instances/inst1/stop"/>
    </actions>
    <public_addresses>
      <address>inst1.larry.fancycloudprovider.com</address>
    </public_addresses>
    <private_addresses>
      <address>inst1.larry.internal</address>
    </private_addresses>
  </instance>
</instances>
```

Quanticate

A PASSION FOR EXCELLENCE

jclouds – jclouds.apache.org

Apache jclouds is an open source multi-cloud toolkit for the Java platform that gives you the freedom to create applications that are portable across clouds while giving you full control to use cloud-specific features.

Quanticate

A PASSION FOR EXCELLENCE

Building out your Solution

Quanticate

A PASSION FOR EXCELLENCE

Tika : tika.apache.org

Text and Metadata extraction

Identify file type, language, encoding

Extracts text as structured XHTML

Consistent Metadata across formats

Java library, CLI and Network Server

SOLR integration

Handles format differences for you

Quanticate

A PASSION FOR EXCELLENCE

UIMA : uima.apache.org

- Unstructured Information analysis
- Lets you build a tool to extract information from unstructured data
- Language Ident, Segmentation, Entities etc
- Components in C++ and Java
- Network enabled – can spread work out across a cluster
- Helped IBM to win Jeopardy!

Quanticate
A PASSION FOR EXCELLENCE

OpenNLP : opennlp.apache.org

- Natural Language Processing
- Various tools for sentence detection, tokenization, tagging, chunking, entity detection etc
- Maximum Entropy, Perception Based M-L
- UIMA likely to be better for a whole-solution
- OpenNLP good when integrating NLP into your own solution

Quanticate
A PASSION FOR EXCELLENCE

cTakes : ctakes.apache.org

Clinical Text Analysis & Knowledge Extraction

Builds on top of UIMA and OpenNLP

Extracts information from free text in electronic medical records (and OCR'd paper)

Identifies named entities from common or custom dictionaries (eg UMLS)

For each, produce attributes, eg subject, mapping, context

Quanticate

A PASSION FOR EXCELLENCE

MINA : mina.apache.org

- Framework for writing scalable, high performance network apps in Java
- TCP and UDP, Client and Server
- Build non blocking, event driven networking code in Java
- MINA also provides pure Java SSH, XMPP, Web and FTP servers

Quanticate

A PASSION FOR EXCELLENCE

Etch : etch.apache.org

Framework for building, producing and consuming network services.

Cross platform, language and transport

Java, C#, C, C++, Go, JS, Python

Produce a formal description of the exchange

1-way, 2-way and realtime communication

Scalable, high performance

Heterogeneous systems

Quanticate

A PASSION FOR EXCELLENCE

Commons : commons.apache.org

Collection of libraries for Java projects

Some historic, many still useful!

Commons CLI – parameters / options

Commons Codec – Base64, Hex, Phonetic

Commons Compress – zip, tar, gz, bz2

Commons Daemon – OS friendly start/stop

Commons Pool – Object pools (db, conn etc)

JMeter : jmeter.apache.org

- Loading testing tool
- Performance test network services
- Defined a series of tasks, execute in parallel
- Talks to Web, SOAP, LDAP, JSM, JDBC etc
- Handy for checking how external resources and systems will hold up, once a big data system start to make heavy use of them!

Quanticate
A PASSION FOR EXCELLENCE

Chemistry : chemistry.apache.org

- Java, Python, .Net, PHP, JS, Android, iOS interface to Content Management Systems
- Implements the OASIS CMIS spec
- Browse, read and write data in your content repositories
- Rich information and structure
- Supported by Alfresco, Microsoft, SAP, Adobe, EMC, OpenText and more

Quanticate
A PASSION FOR EXCELLENCE

ManifoldCF : manifoldcf.apache.org

- Framework for content (mostly text) extraction from content repositories
- Aimed at indexing solutions, eg SOLR
- Connectors for reading and writing
- Simpler than Chemistry, but also works for CIFS, file systems, RSS etc
- Extract from SharePoint, FileNet, Documentum, LiveLink etc

Quanticate
A PASSION FOR EXCELLENCE

Questions?

Quanticate

A PASSION FOR EXCELLENCE

Thanks!

Twitter - @Gagravarr

Email – nick.burch@quanticate.com

The Apache Software Foundation:
http://www.apache.org/

Apache projects list:
http://projects.apache.org/

Quanticate

A PASSION FOR EXCELLENCE