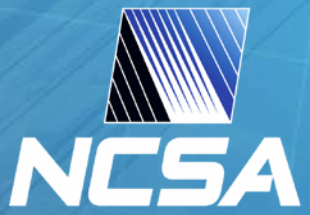
Medici 2 Tutorial

Luigi Marini

October 22nd, 2013



National Center for Supercomputing Applications
University of Illinois at Urbana–Champaign

Goals

- Provide some context for why a rewrite (version 2)
- Get everyone comfortable with the v2 code base
  - Be able to contribute
- Mostly hands on

- Not goals:
  - Discuss version 1 in depth
  - Discuss every version 2 feature in depth

Agenda

- Short Medici past / present / future
- Code review of version 2
- Writing new extractors
- Writing new previewers

~45 mins each

Setup Development Environment

- Java (1.6 or 1.7)
- Scala sbt 0.13.0
- http://www.scala-sbt.org/
- Scala IDE 3.0 (as plugin on existing eclipse or standalone)
- http://scala-ide.org/
- Python virtualenv
- https://pypi.python.org/pypi/virtualenv
- RabbitMQ Server
- http://www.rabbitmq.com/download.html
- MongoDB
- http://www.mongodb.org/downloads

BRIEF MEDICI
PAST / PRESENT / FUTURE

When we started…

- “Manage large collections of multimedia research artifacts”
- Focus on image data
- Was using Tupelo and RDF to store everything
- Development split between web and desktop

What parts were successful?

- Decentralized cloud storage
  - Install your own instance and maintain for your community
- Flexible metadata support
  - No predefined schemas/ontologies
- Framework for plugging domain specific
  - Extractors
    - Dig into the files for information
  - Previewers
    - Visualize information on the web
- Support for discovery of new content
  - Text-based search, social annotation

Timeline

• v.1	Development started	November 2009
• v.1	First public release	June 2010
• v.1.1	New features	November 2010
• v.1.2	New features	February 2012
• -	Review	Summer 2012
• v.2	Development started	November 2012
• v.1.3	SEAD	November 2013
• v.2	Next release	December 2013

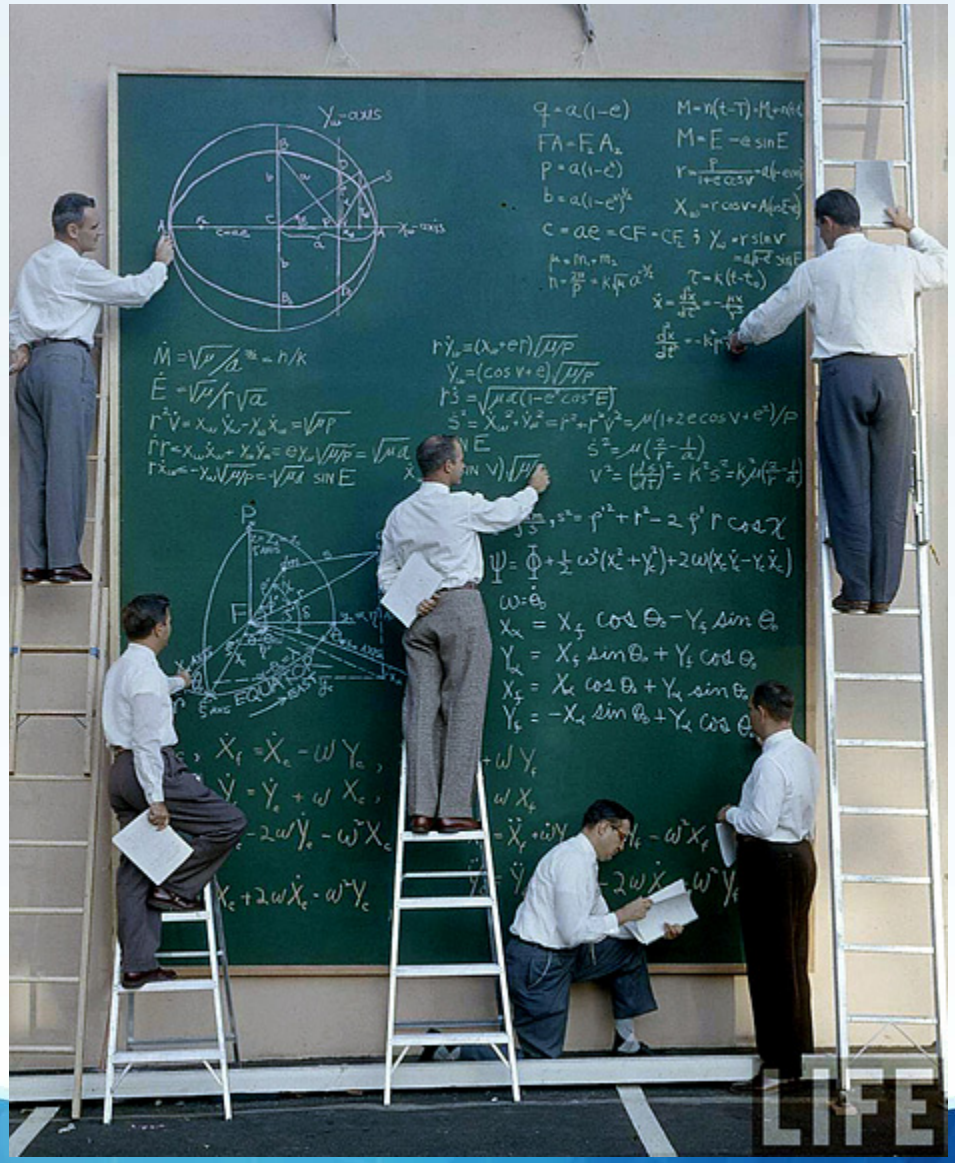
The Review of Summer 2012…

- It was a hot and humid Summer in the prairie….

The Review of Summer 2012…

- We hit a ceiling on the number of RDF triples we could store
- Tupelo (Java RDF library) was incredibly difficult to debug
  - Especially since the main developer had left
- Public dataset could not be index by search engines
  - Because it was all javascript
- Using plain Javascript libraries with GWT was always a challenge and a time sync
- State stored in the servlet session server side made it hard to scale horizontally

Back to the drawing board!



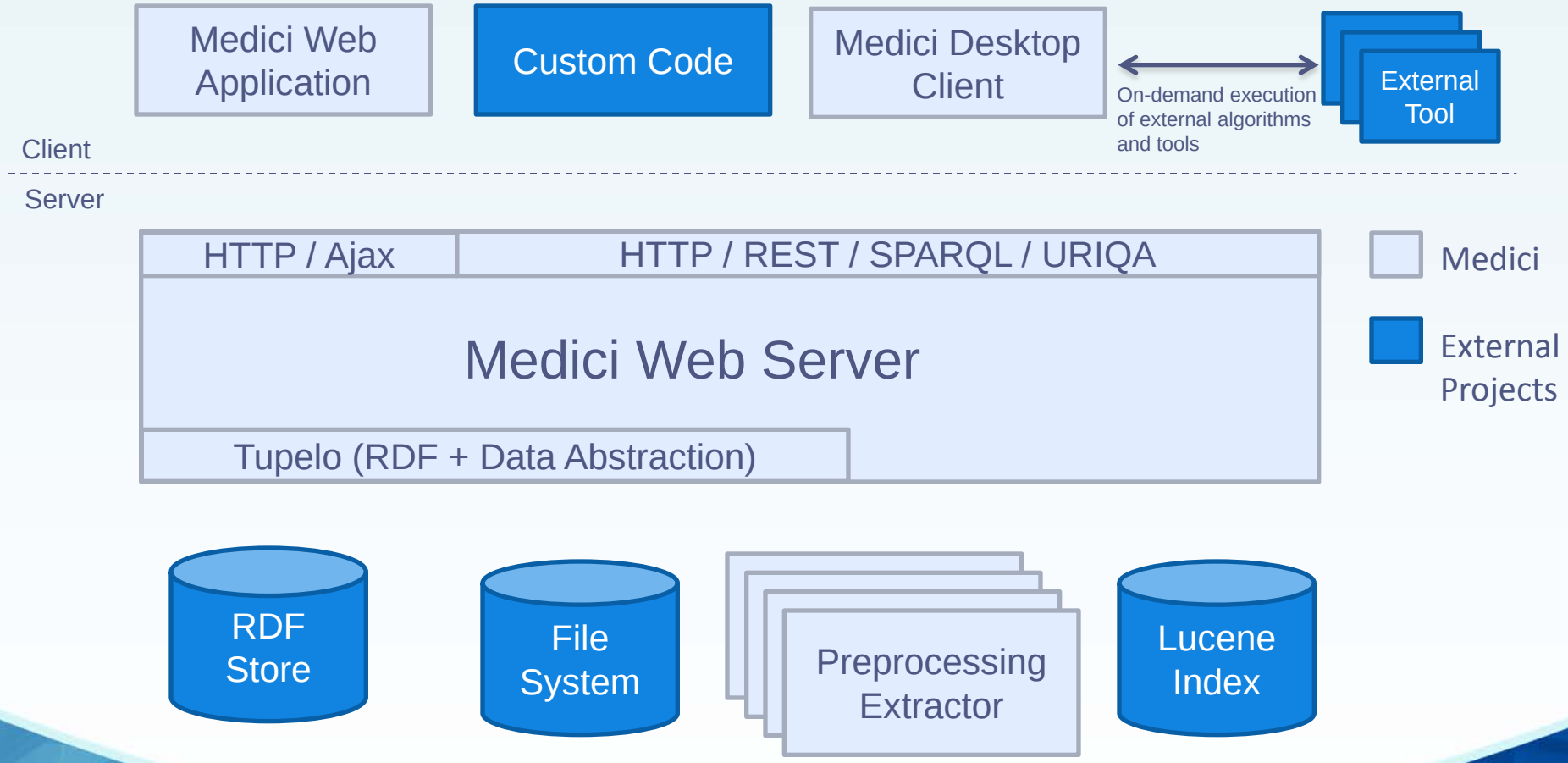
Priorities

- Scalable
- Maintainable
- Extensible

Improve the design

- Add “Projects” to moderate access and provide groups
- Provide “activity streams” to keep the researcher in the loop
- Improve retrieval
  - Signals
  - Multimedia
- Enable on demand processing

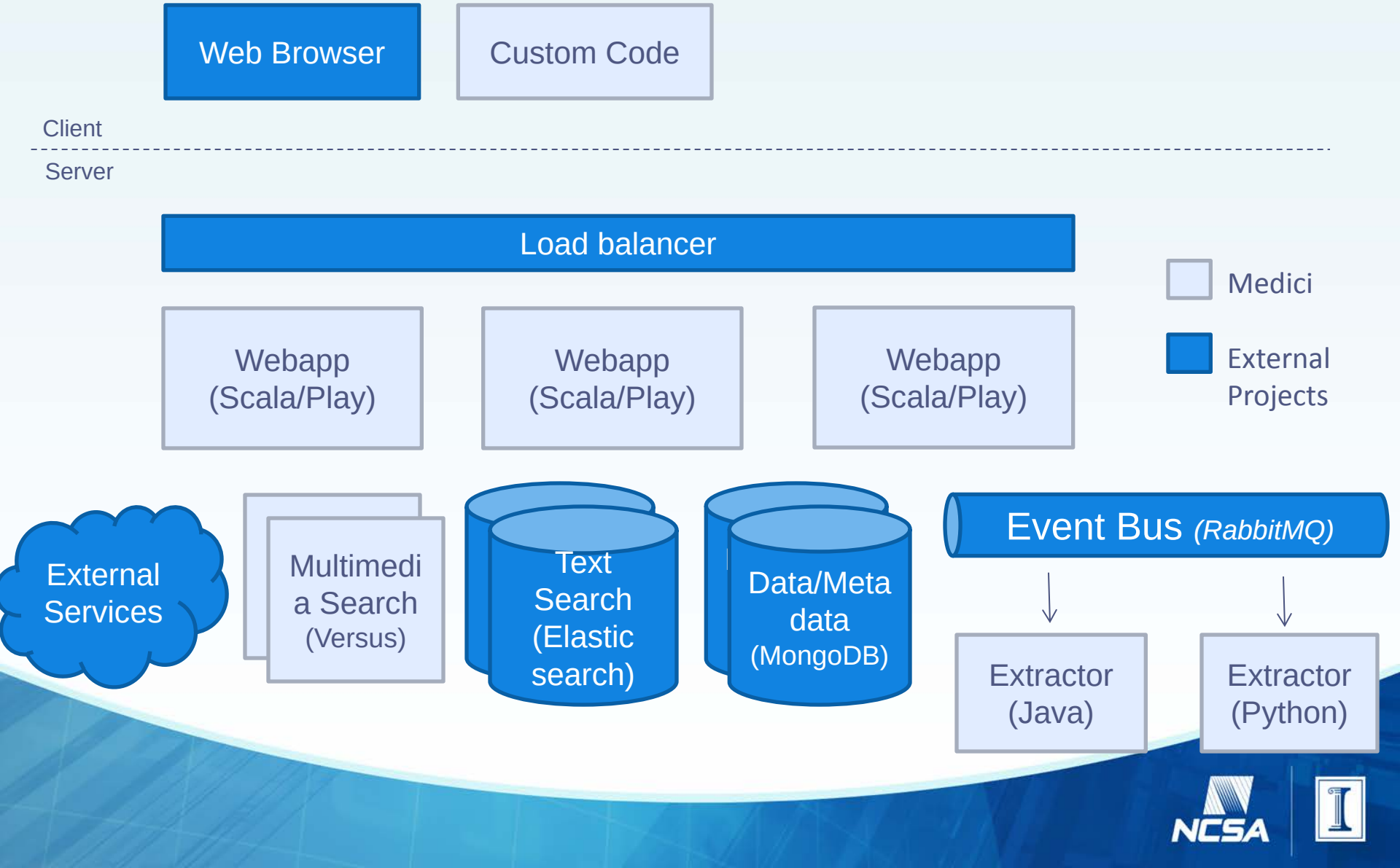
Architecture (v.1)



Technologies (v.1)

- Web application
  - Google Web Toolkit
  - Java Servlets
  - Plain Javascript
  - Viewers: Flash, Java Applet, HTML, etc.
  - Apache Lucene
  - Mysql
- Extraction Service
  - Eclipse RCP (Java)
  - Large collection of external applications
- Desktop Client
  - Eclipse RCP (Java)
  - Cyberintegrator Workflow Management System

Architecture (v.2)



Scala

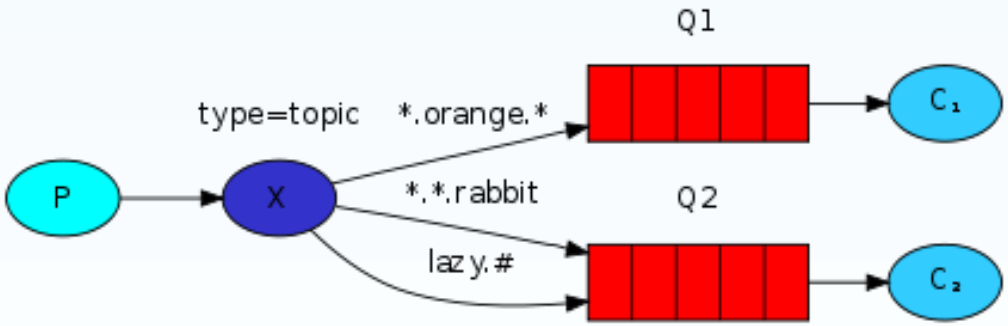
- Build on the JVM
  - Can use Java libraries
- Object-Oriented programming meets Functional programming
- Actor model for concurrency
- Of Twitter fame

Play

- Scala and Java versions
- Lightweight and modular
- Akka.io
  - High performance concurrent and distributed application
  - Actor based programming (of Erlang fame)
- Netty.io
  - Asynchronous event-driven network application framework

RabbitMQ

- Publish Subscribe message bus
- Written in Erlang
- AMPQ standard
- Drivers in most languages
- Of Reddit fame



MongoDB

- Document-oriented storage
- JSON
- GridFS for large binary files
- Sharding and replication out of the box
- Drivers in many languages
- Easy to install and get started with
- Difficult to manage large cluster

Elasticsearch

- Built on Lucene
- Sophisticated RESTful services API using JSON
- Distributed out of the box
- Proven track record

New features in version 2

- Introducing spaces
  - Group based access control
- File versioning
  - Keeping generic provenance trail
- Multiple files in a dataset
  - Explicit instead of implicit zip files
- Multimedia search
  - Find similar images, videos, audio
- Better relationship management between datasets

Future?

- On demand execution of tools from the web
- Linked data services (RDF is back!)
- Support for different backends
  - Cassandra
  - HDFS
- Real geospatial support using PostGIS

Me-di-ci

- Was this the same thing as what we started with?

Open Research Content

CODE REVIEW

https://opensource.ncsa.illinois.edu/stash/projects/MED/repos/medici-play

New Previewer & Extractor

- http://bl.ocks.org/mbostock/3884955