Apache Big Data: 1 petabyte per second

SUSAN MALAIKA / SEPTEMBER 30, 2015 / 0 COMMENTS

The Apache Big Data Conference:

The first Apache Big Data Conference is taking place in Budapest. There are more than 300 enthusiastic contributors and consumers of Apache Big Data projects attending the conference. You can see associated tweets #apachebigdata

The conference location is sumptuous, the schedule is intense, the sessions are energizing, the conversation is stimulating, and the food is delicious. A thoughtful touch at the conference, is they have single sheets with a summary of all sessions and locations for one day – handy when trying to determine where to go next. Popular topics at the conference are streaming, ubiquitous SQL and tabular interfaces, performance comparisons, streamlining builds – tests – deploys, data science, statistics, machine learning and getting started with Apache code, and how to become an Apache contributor. There were plenty of discussions on Apache CouchDB, Kafka, and Flink and often sessions combined 2-5 projects. There were sessions on the ODPi which I describe separately.

During the opening keynotes, we heard from Shane Curcuru that Apache has 167 active projects with over 5,000 committers. The committers have varied backgrounds, and include a pilot for a national airline. On one day Apache serves 2.2 million distinct users

in 206 countries. Throughout the conference we could see who was committing Apache code by country on a screen in the expo area.

The beautiful ballroom, with portraits of eminent citizens, where Apache Big Data keynote sessions are held

Dirk Düllmann from CERN told us about huge volumes of data such as generating 1 petabyte / second from a Higgs boson collider. Amongst other things, CERN uses Hadoop to analyze the data that describes their machine clusters on the network – they need to use Hadoop because of the size of the network and amount of metadata!

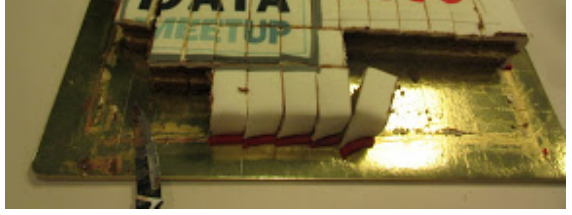
At the Apache Barcamp session (where the agenda is created on the fly) we reviewed the newly designed Apache pages including https://projects.apache.org/ and http://status.apache.org. On the project pages you can see the 30 Apache projects that are classified as Big Data. On the status pages, on the final day of the Big Data conference, we observed that HBase was the most active committer project, and Spark was the third most active project on the mailing list, on that day. After reviewing the mentoring pages at Apache, we prepared an outline of a "getting started in open source" guide.

Budapest Big Data Meetup

I attended a breakout presentation that A Tale of Two Graphs : Property Graphs and RDF given by Andy Seaborne and Paolo Castagna as I had given a similar presentation at an internal graph summit that my colleagues Kelvin Lawrence and Marie Wallace hosted recently – although my session was particularly focused on open data. I was delighted that the conclusions were comparable: that there are different use cases for each type of graph.

The Budapest Big Data Meetup

On the first evening some of us at the Apache Big Data conference attended the local Budapest Big Data meetup group nearby – organized by Bence Arato. There were about 50 attendees including the friendly local community. Budapest has been a wonderful source of successful technology startups, and I was fortunate to meet some of them. There were engaging presentations on the Internet of Anything, SQL and streaming interfaces on Big Data, and Data Science. One of the talks featured Thomas Bayes' grave! We were lucky: The Budapest meetup group leadership were celebrating their 1000th member, so we all had chocolate cake.



Budapest Big Data Meetup Cake

Share this:

3

Leave a comment

Name *

Email *

Comment

Submit Comment

Notify me of follow-up comments by email.

Notify me of new posts by email.