SPARK SUMMIT EAST 2017 – NOTES

8. Februar (37/63)

Keynote (5/5)

WHAT TO EXPECT FOR BIG DATA AND APACHE SPARK IN 2017

9:00 AM – 9:20 AM
Matei Zaharia from Databricks
link
video
Talks about the plans. Python, R will get more attention.
Spark is used in Facebook (they are the Hive producers), CapitalOne for Credit Fraud prevention.
Languages in Spark in 2016: Scala 65%, Java 29%, Python 62%, R 20%. SQL 95%

USING APACHE SPARK FOR INTELLIGENT SERVICES

9:20 AM – 9:40 AM
Alexis Roos from Salesforce
link

video
Data acquisition: Kafka
Processing: Streaming, NLP, Deep learning, Graph
Insights:
Suggestions: suggest activities
Enriched mail/conversation sent into Kafka queue

Practical talk with a demo and code – Inbox Sales email demo.
How to optimize CRM usage with big data technologies and improve recommendation.

PRODUCTION-READY STRUCTURED STREAMING

9:40 AM – 9:55 AM
Michael Armbrust from Databricks
link
Practical talk, showing the usage in databricks notebook.
Time series graphs with real time data, very cool.
Hist data + streaming data -> combined into one stream for analysis. S3 & Kafka -> Parquet for Ad Hoc Queries.
All this is available on Databricks blog!

SCALING GENETIC DATA ANALYSIS WITH APACHE SPARK

9:55 AM – 10:15 AM
Cotton Seed from Broad Institute of MIT and Harvard
link
Talks about Hail – open source framework for genetic data. Hail is available on Databricks for testing.
https://hail.is

RISELAB: ENABLING INTELLIGENT REAL-TIME DECISIONS

10:15 AM – 10:30 AM
Ion Stoica from UC Berkeley AMP/RISE Lab & Databricks
link
slides pdf
RiseLab follows AmpLab
Project started in January 17. Focuses on real time decisions on live data
Improving Apache Spark:
Project Drizzle – decrease latency if structured streaming and ML algorithms. Drizzle will be embedded in Spark
Project Opaque – full data encryption, authentication and verification.

Spark Ecosystem - Ballroom A (9/9)

NEW DIRECTIONS IN PYSPARK FOR TIME SERIES ANALYSIS

11:00 AM – 11:30 AM
David Palaitis from Two Sigma
link
video
Two Sigma has developed Flint - a library for time series that works on top of Spark. It is available on github and the APS are Scala and Python.
Storing data in parquet in HDFS and then loading them to Spark is the dominant practice.

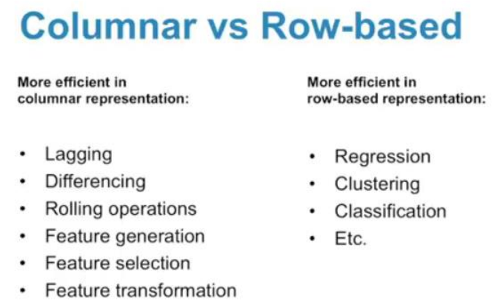
TIME SERIES ANALYTICS WITH SPARK

11:40 AM – 12:10 PM

Simon Ouellette from Faimdata

[link]

[video]



Presenting a time-series library in Spark.

LESSONS LEARNED FROM DOCKERIZING SPARK WORKLOADS

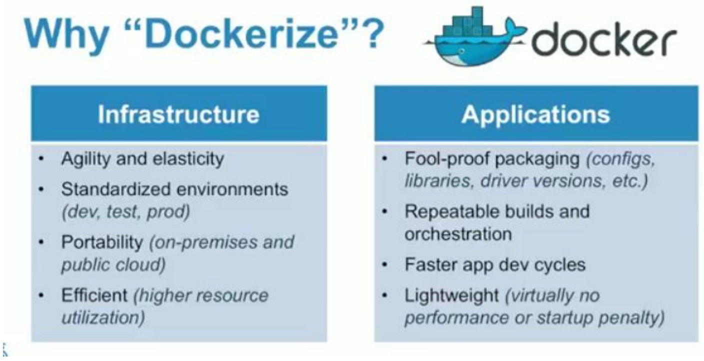
12:20 PM – 12:50 PM

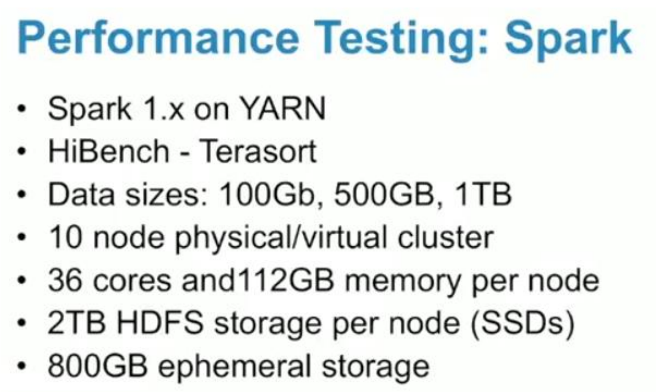
Tom Phelan from BlueData

[link]

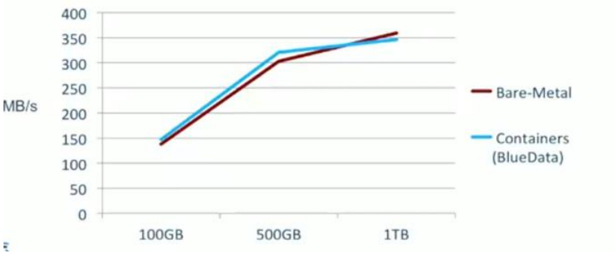
[video]

Talk about Docker, Spark on Docker, Zeppelin on Docker. Very informative if one goes into Docker.





Spark on Docker: Performance



Spark on Docker: Key Takeaways

- All apps can be "Dockerized", including Spark
  - Docker containers enable a more flexible and agile deployment model
  - Faster app dev cycles for Spark app developers, data scientists, & engineers
  - Enables DevOps for data science teams

Spark on Docker: Key Takeaways

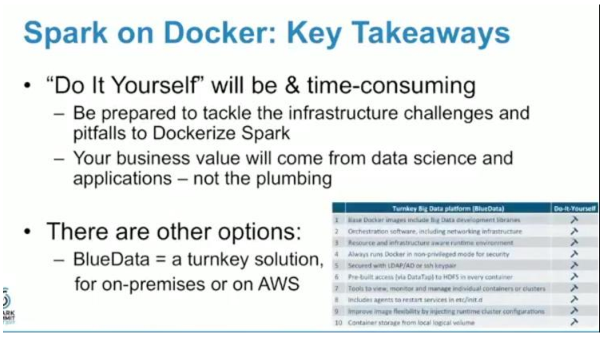
- Deployment requirements:
  - Docker base images include all needed Spark libraries and jar files
  - Container orchestration, including networking and storage
  - Resource-aware runtime environment, including CPU and RAM

Spark on Docker: Key Takeaways

- Data scientist considerations:
  - Access to data with full fidelity
  - Access to data processing and modeling tools
  - Ability to run, rerun, and scale analysis
  - Ability to compare and contrast various techniques
  - Ability to deploy & integrate enterprise-ready solution

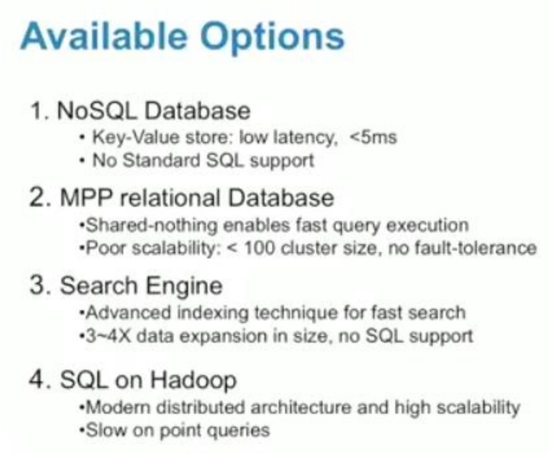
Spark on Docker: Key Takeaways

- Enterprise deployment challenges:
  - Access to container secured with ssh keypair or PAM module (LDAP/AD)
  - Fast access to external storage
  - Management agents in Docker images
  - Runtime injection of resource and configuration information

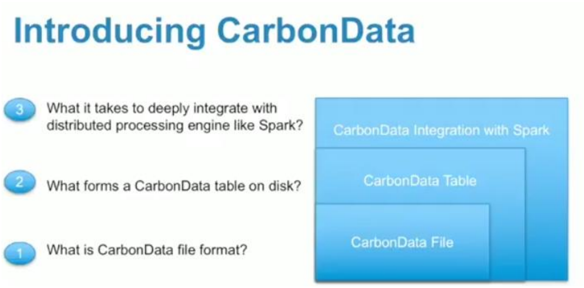


APACHE CARBONDATA: AN INDEXED COLUMNAR FILE FORMAT FOR INTERACTIVE QUERY WITH SPARK SQL

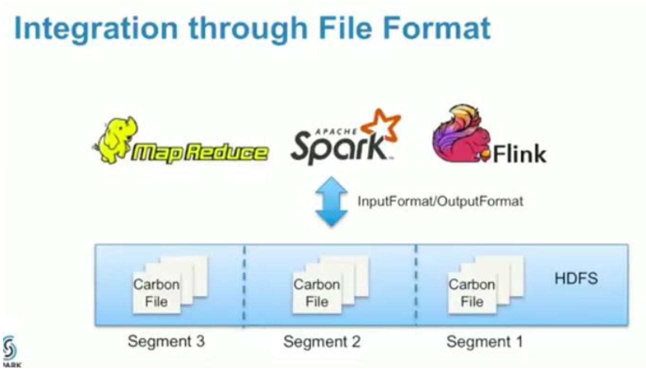
Jacky Li from Huawei Technologies & Jihong Ma from Huawei
2:00 PM – 2:30 PM
link
video
Practical talk, detailed explanation of the file format with examples.
Storage options on market:



They developed their own file format system – CarbonData – Apache incubator project since june 2016.



It is indexed columnar file format.
Carbon file can be stored in HDFS.
CarbonData is used for interactive data analysis with SparkSQL (Spark 2.1) on CarbonData tables.



Tests show CarbonData was faster than Parquet.

BUILDING REALTIME DATA PIPELINES WITH KAFKA CONNECT AND SPARK STREAMING

2:40 PM – 3:10 PM
Ewen Cheslack-Postava from Confluent
link
video
Kafka Connect is large-scale data streaming import/export for Kafka.
Many Kafka connectors available on Confluent website.
Very useful talk when going into Kafka development.
No practical examples, but a well described architecture.

BUILDING A DATASET SEARCH ENGINE WITH SPARK AND ELASTICSEARCH

3:20 PM – 3:50 PM
Oscar Castañeda-Villagrán from Xoom, a PayPal service
link
video
Metadata extraction is the core of the talk.
How to run ElasticSearch in Spark cluster is explained. Code is shown and detailed explanation is given. Extracting metadata from the DataSets is done.
ES snapshots are saved to Amazon S3. Demo is shown.

TEACHING APACHE SPARK CLUSTERS TO MANAGE THEIR WORKERS ELASTICALLY

4:20 PM – 4:50 PM
Erik Erlandson & Trevor Mckay from Red Hat
link
video
Containerizing Spark.
Explaining the benefits of using containers.
They developed Oshinko – tool for creating cluster with default settings. Scale and delete is simple.

APACHE TOREE: A JUPYTER KERNEL FOR SPARK

5:00 PM – 5:30 PM
Marius van Niekerk from Maxpoint
link
video
Toree is an implementation of the Jupyter Kernel Protocol.
Talk about the Toree project and embedding it with Jupyter.

SECURED (KERBEROS-BASED) SPARK NOTEBOOK FOR DATA SCIENCE

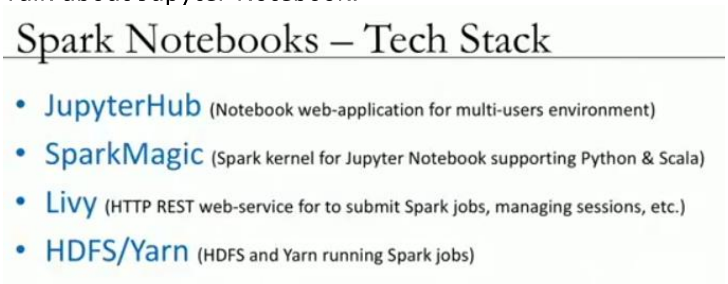
5:40 PM – 6:10 PM

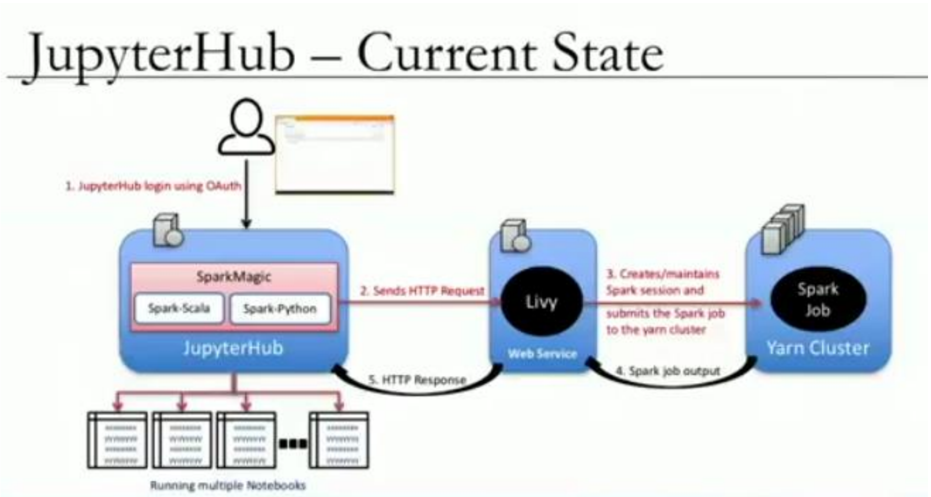
Joy Chakraborty from Bloomberg

link

video

Talk about Jupyter Notebook.





Explanation of Jupyter & Spark with Kerberos.

Developer - Ballroom B (9/9)

PROCESSING TERABYTE-SCALE GENOMICS DATASETS WITH ADAM

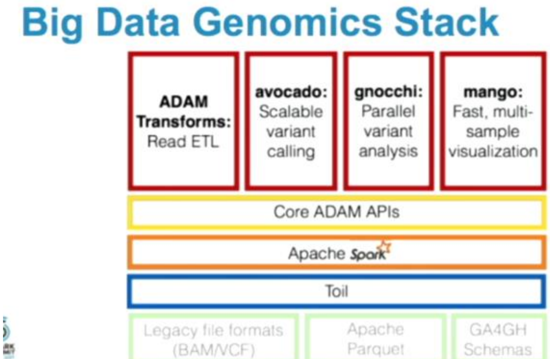
11:00 AM – 11:30 AM

Frank Austin Nothaft from UC Berkeley

link

video

In charge of Adam project. ADAM is open source, distributed library for genomic analysis. Spark Scala API is available. Java and Python are coming. Batch analysis available, exploratory analysis is one of the goals.



ADAM produces statistically equivalent results to the GATK, only faster. Read preprocessing is 30x faster and 3x cheaper, end-to-end is 4x faster and 3,5x cheaper.

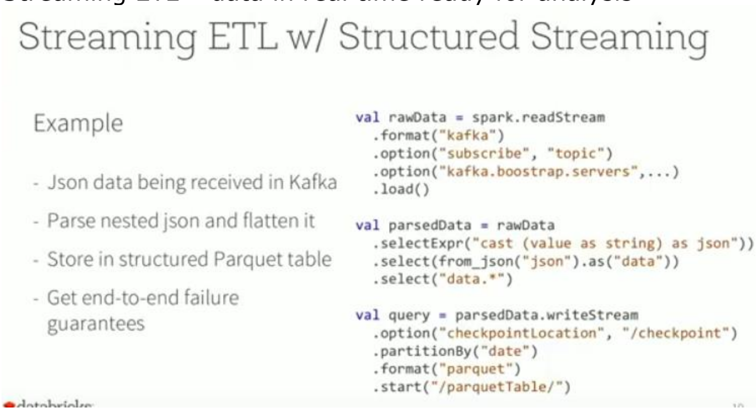
Whole project is on github and 2 notebooks in Databricks for testing.

MAKING STRUCTURED STREAMING READY FOR PRODUCTION - UPDATES AND FUTURE DIRECTIONS

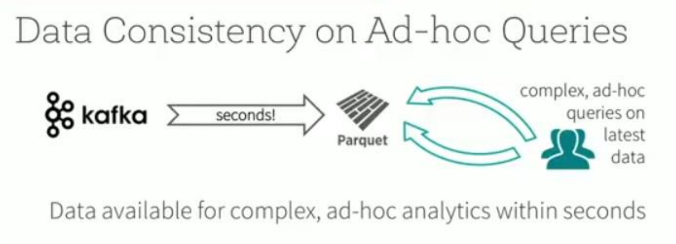
11:40 AM – 12:10 PM
Tathagata Das from Databricks
[link]
[video]
Talk about structured streaming.
Treat streams as unbounded tables – append new rows to table.

->

Fault-tolerance with checkpointing. – metadata of current batch is stored in a write ahead log in HDFS/S3.
Traditional RTL – hours before data is analyzed
Streaming ETL – data in real time ready for analysis

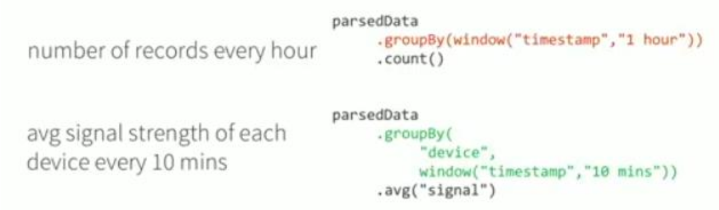


Partition data per date so that queries on time are faster.



[link to blog on this]

Windowing is another type of grouping in structured streaming.



Currently Scala and Java only – Spark 2.1
Many more updates on streaming in Spark 2.2

Comparison with Other Engines

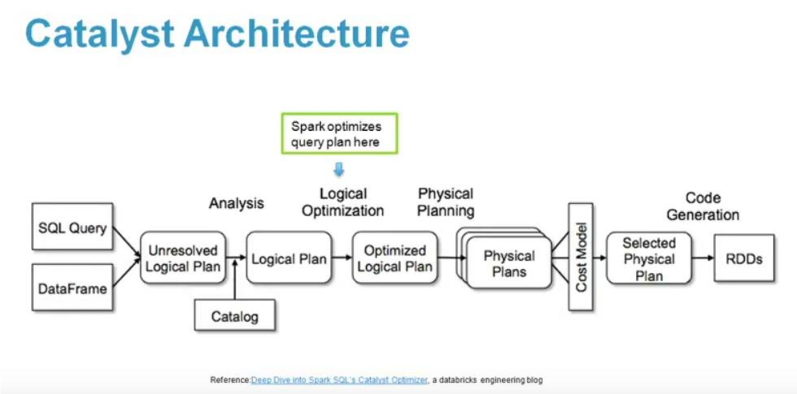
Property	Structured Streaming	Spark Streaming	Apache Storm	Apache Flink	Kafka Streams	Google Dataflow
Streaming API	incrementalize batch queries	integrates with batch	separate from batch	separate from batch	separate from batch	integrates with batch
Prefix Integrity Guarantee	✔	✔	✘	✘	✘	✘
Internal Processing	exactly once	exactly once	at least once	exactly once	at least once	exactly once
Transactional Sources/Sinks	✔	some	some	some	✘	✘
Interactive Queries	✔	✔	✘	✘	✘	✘
Joins with Static Data	✔	✔	✘	✘	✘	✘

databricks Read the blog to understand this table

Deeper introduction to Streaming Spark. Very concrete and useful. Kafka reference and usage in streaming with Spark.

COST-BASED OPTIMIZER FRAMEWORK FOR SPARK SQL

12:20 PM – 12:50 PM
Ron Hu and Zhenhua Wang from Huawei Technologies
[link]
[video]
Very difficult to understand the speaker.
Huawei contributes to Spark community with CBO.



CBO in Spark SQL was done in Physical Plans and Cost Model.
Spark CBO coming in Spark 2.2.

OPTIMIZING APACHE SPARK SQL JOINS

2:00 PM – 2:30 PM
Vida Ha from Databricks
[link]
[video]
Very informative talk about joins in Spark. Explaining 4-5 joins and how to use them.
Shuffle Hash Join & Broadcast Hash Join – two most common and used.
SHJ goes back to MapReduce fundamentals:
1) Map – creates output key
2) The key is used in shuffle
3) Reduce phase
Check Spark UI for task level detail if you want to detect shuffle problems.

Parquet is recommended for Spark SQL.

Cartesian join is also mentioned and explained.

Theta join.

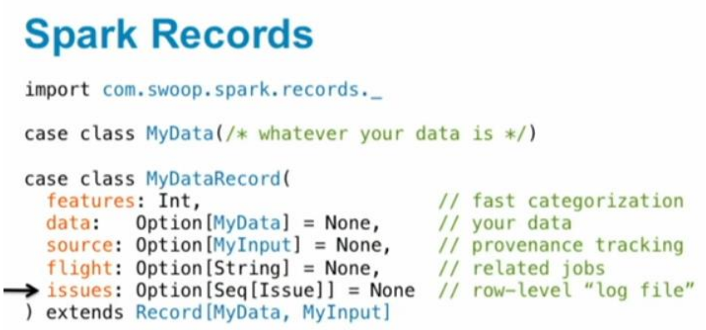
From Q&A: Spark spills to the disk when needed. Biggest partition needs to fit the biggest instance.

THE JOY OF NESTED TYPES WITH SPARK

2:40 PM – 3:10 PM
Ted Malaska from Blizzard
[link]
[video]
Very practical presentation.
It is possible to debug Spark code, as long as you have a JVM (he is showing it on his computer)
Malaska's git repositories are useful for learning more on this topic.

BULLETPROOF JOBS: PATTERNS FOR LARGE-SCALE SPARK PROCESSING

3:20 PM – 3:50 PM
Sim Simeonov from Swoop
[link]
[video]
They build clients on Spark.
Spark vs. magic. Magic is a client on top of Spark.
Spark Records – available on github.
Root-cause analysis and how to make your life easier with Spark Records.
Practical talk, with example in Databricks Notebook.



[website]
[github]

WHAT NO ONE TELLS YOU ABOUT WRITING A STREAMING APP

4:20 PM – 4:50 PM
Ted Malaska from Blizzard
[link]
[video]

He was critical against Spark Streaming. He claimed there is no need for spark streaming. He is using Spark with forever loop. Ted mentioned that all Lucene products (Solr, Elastic) are expensive- price and resource. Ted: Lambda is 10-year-old architecture, outdated.

HORIZONTALLY SCALABLE RELATIONAL DATABASES WITH SPARK

5:00 PM – 5:30 PM
[link]
A lot of talk about Postgresql. A bit unclear when it comes to delivering the message.

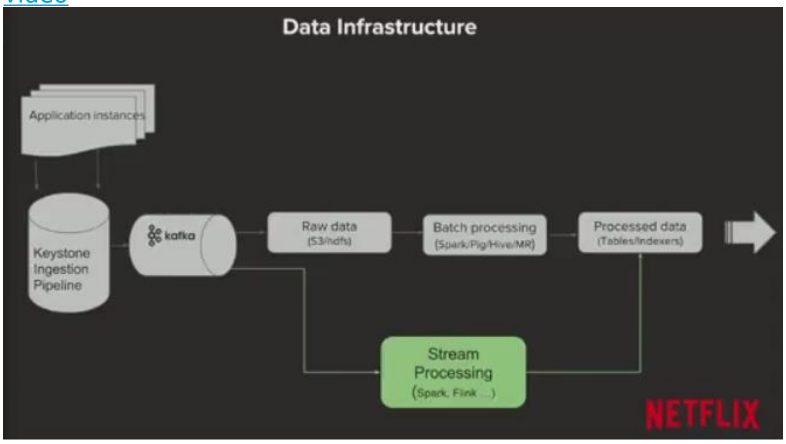
SPARK AND OBJECT STORES —WHAT YOU NEED TO KNOW

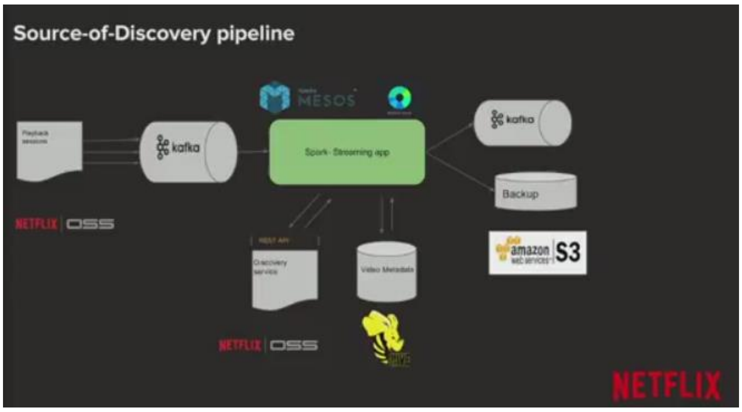
5:40 PM – 6:10 PM
Steve Loughran from Hortonworks
[link]
Steve talks about experience with object stores in AWS and cooperation with Cloudera on S3guard.
Near future: streaming on cloud
He recommends to use S3A.
How to connect to S3A from Spark. Spark Streaming in cloud - han vari tvil om dette fungerer. S3A in DataFrames - code examples. Do not commit data diretly to S3A, save to HDFS first and data uwant to perserve, commit it to S3A! Hadoop 2.8 offers better S3A performance. Example: Azure storage and Streaming.

Spark experience and Use cases - Ballroom C (9/9)

GOING REAL-TIME: CREATING FREQUENTLY-UPDATING DATASETS FOR PERSONALIZATION

11:00 AM – 11:30 AM
Shriya Arora from Netflix
[link]
[video]





Talk about Spark Streaming, explained, how to optimize it, what to be aware of, performance tuning…
Interesting to see how, what Netflix is doing on Spark Streaming and challenges they face.

SPARK FOR BEHAVIORAL ANALYTICS RESEARCH

11:40 AM – 12:10 PM
John Wu from Berkeley Lab
link
video
Work among data scientists with behavioral economists
Research on energy policy.
No Spark talk, just talk about electricity consumption and analysis of it.

SPARK AS THE GATEWAY DRUG TO TYPED FUNCTIONAL PROGRAMMING

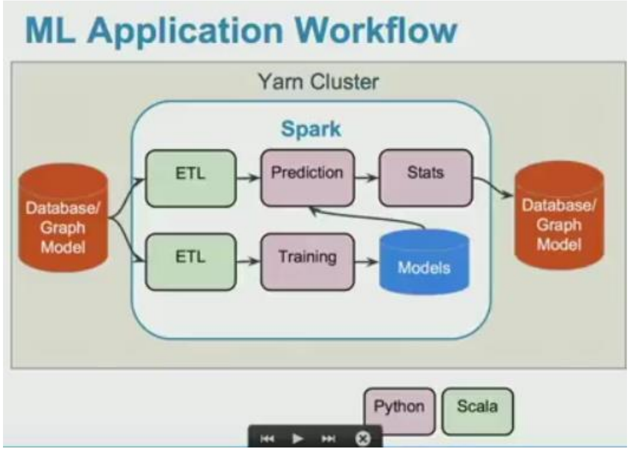
12:20 PM – 12:50 PM
Jeffrey Smith & Rohan Aletty from x.ai
link
video
Spark & Scala & Typed Functional Programming
Focus of the talk: How to build TFP knowledge in your team? Very structured talk about how to start with Scala in Spark.
Monads in Scala: container on which you can run some operations on values in the container. List, Set, Option are Monads in Scala.

EXPLORING SPARK FOR SCALABLE METAGENOMICS ANALYSIS

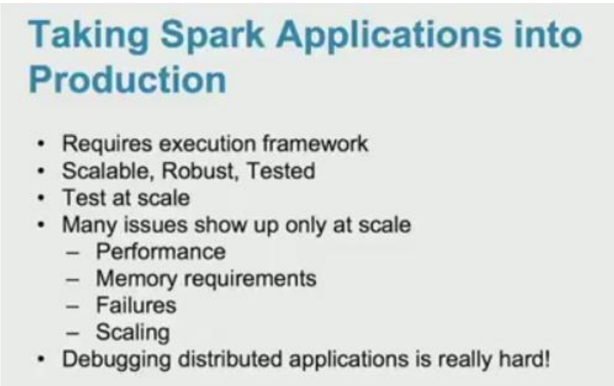
2:00 PM – 2:30 PM
Zhong Wang from DOE Joint Genome Institute
link
video
Talk about genomics.
In 2013 developed Hadoop/BioPig – runs on AWS. Had many challenges (slow, expensive). Moved to Spark.
Standalone Spark on single large memory server. AWS Elastic Map Reduce.
Not so much Spark talk, more like high-level talk about the work they do and why they went with Spark.

MONITORING THE DYNAMIC RESOURCE USAGE OF SCALA AND PYTHON SPARK JOBS IN YARN

2:40 PM – 3:10 PM
Ed Barnes & Ruslan Vaulin from Sqrrl Data
link
video
They have developed a tool for monitoring and debugging Spark Jobs. Django/Apache server with MySql.



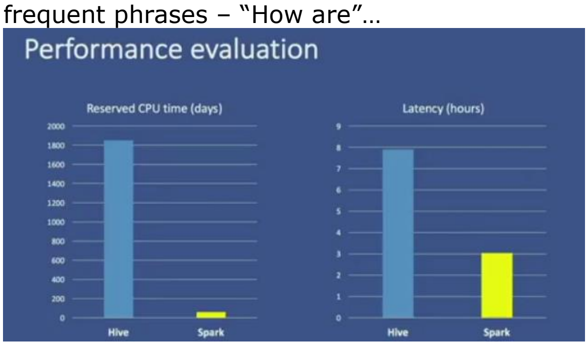
Data is in a distributed database.



EXPERIENCES WITH SPARK'S RDD APIS FOR COMPLEX, CUSTOM APPLICATIONS

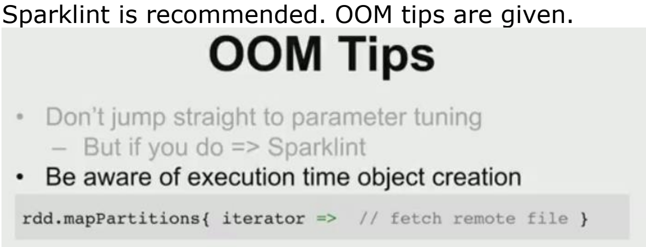
3:20 PM – 3:50 PM
Tejas Patil from Facebook
link
video
Natural Language Processing example is used to talk about moving from Hive to Spark and the challenges they faced in Spark (Data skew).
Hive allows you to run UDF or run a process - TRANSFORM (data is serialized, pushed into the process, processed and Hive takes the results back)
Previous solution for NLP was in Hive – quite some challenges.

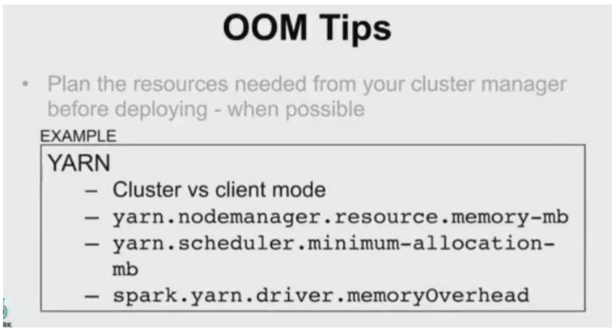
They moved from Hive to Spark.
Operator *pipe()* allows you to run any code from Spark (f. ex. C++ code. It is like TRANSFORM in Hive that they used in previous solution).
Data skew was the main challenge and root to many errors. Data skew happened because of frequent phrases – "How are"…

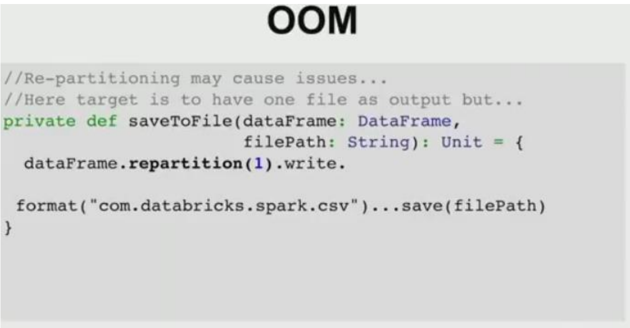


In this use-case they partition by last two words – this is because of the algorithm they use.

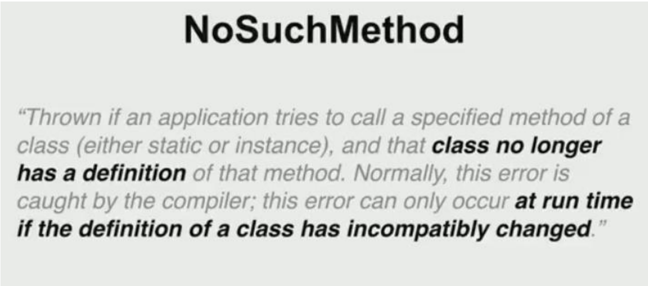
PROBLEM SOLVING RECIPES LEARNED FROM SUPPORTING SPARK
4:20 PM – 4:50 PM
Justin Pihony and Stavros Kontopoulos from Lightbend
[link]
[video]
- Out of Memory (OOM) problem.
Spark.memory.fraction vs spark.memory.storageFraction
Sparklint is recommended. OOM tips are given.

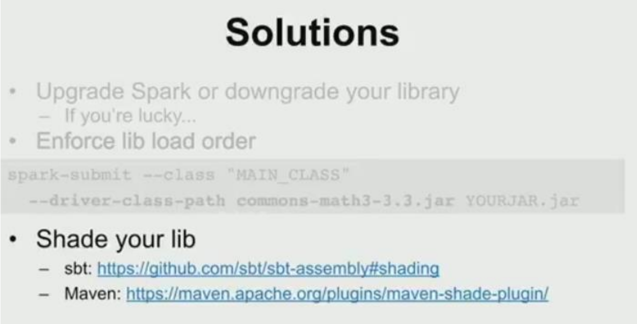






- NoSuchMethod

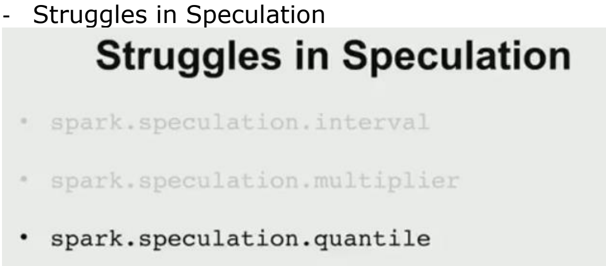




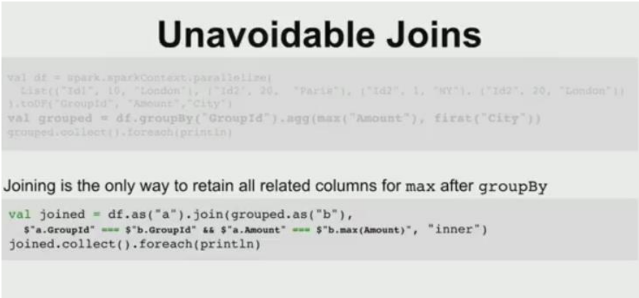
- Perplexities of Size

Parallelism – 2 to 3 times the amount of cores (?)

- Struggles in Speculation



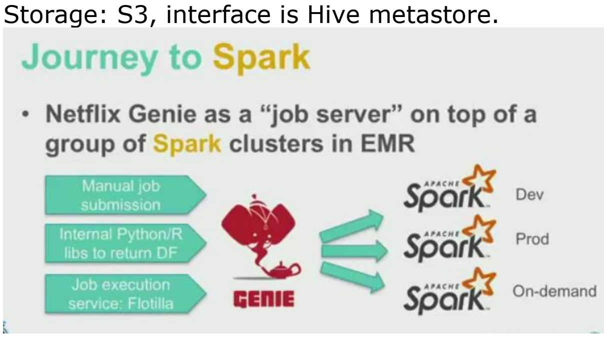
- Slow Joins



- Handling S3 without hanging

MIGRATING FROM REDSHIFT TO SPARK AT STITCH FIX

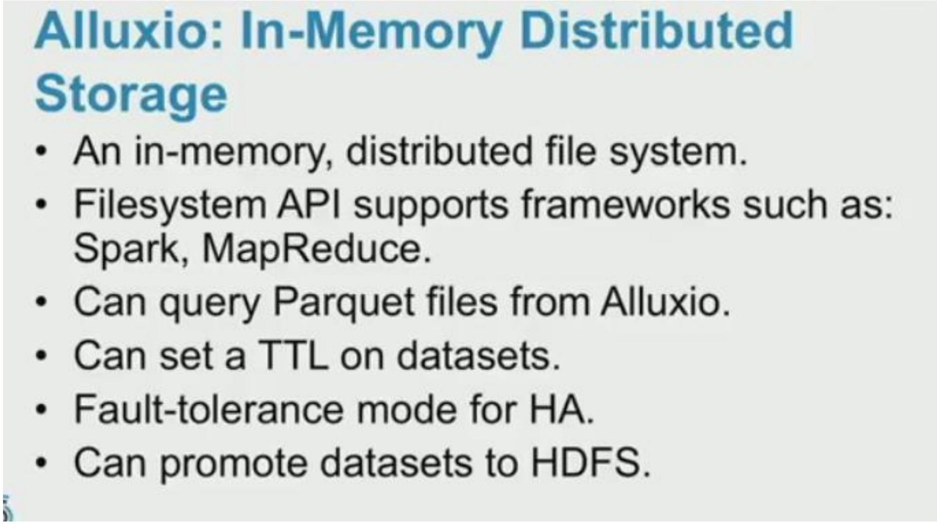
5:00 PM – 5:30 PM
Sky Yin from Stitchfix
[link]
[video]
Cloth recommendation business – they recommend clothes by using human stylists and algorithms.
80 people in data team, biggest table adds 500-800M rows per day.
Heavy Python and R users, deployed through Docker.
They wanted to divide storage and computation.
Storage: S3, interface is Hive metastore.



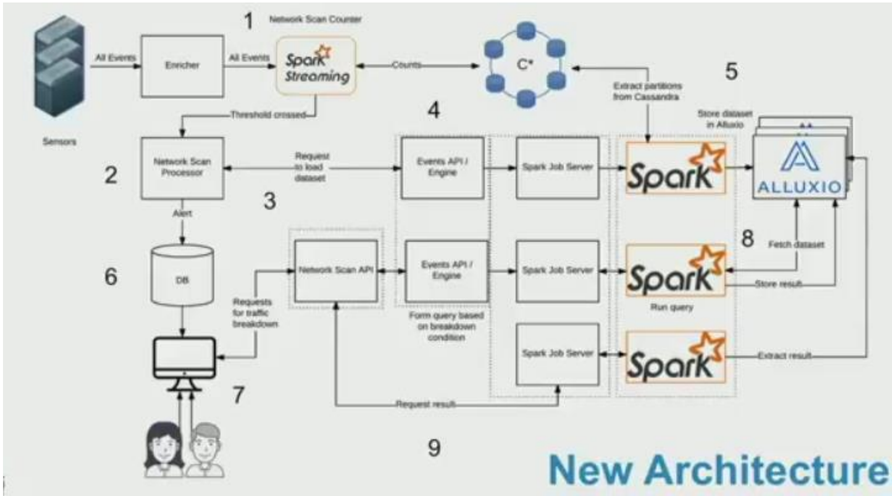
Genie is an opensource Netflix solution. Driver node runs inside of Genie.
Very concrete talk about Spark, what to focus on, best practices…

FIGHTING CYBERCRIME: A JOINT TASK FORCE OF REAL-TIME DATA AND HUMAN ANALYTICS

5:40 PM – 6:10 PM
William Callaghan from eSentire
[link]
[video]
Storage is Cassandra. Data is modelled based on queries, not relations.
Spark Job Server for submitting jobs via REST API.



Use Alluxio, files are in parquet format. Code examples are shown and explained in the presentation.

New architecture:


One Spark job pulls data from Cassandra into Alluxio. Another Spark job takes the data in Alluxio and analyzes it.

Data Science - ROOM 302/304 (2/9)

TUNING AND MONITORING DEEP LEARNING ON APACHE SPARK

2:00 PM – 2:30 PM
Tim Hunter from Databricks
[link]
[slides pdf]
2016: year of emerging Spark & Deep Learning & GPU.
Many DL frameworks with Spark bindings (TensorFlow…). Many are running natively on Spark (MLlib…).
Databricks perspective:

Spark vendor on public cloud.
Provides GPU instances
Helps with compute-intensive workloads

Paper: Hidden technical debt in ML algorithms, Sculley NIPS 2016 - (where ML fits, what needs to be done in the pipeline)
Talks about DL in the pipeline.
Patterns in DL development:

- data stored outside of Spark
- DL transforms: data stored in DF/RDD
- multiple passes over data

PySpark is popular when using GPU. Many DL packages have Python interface.Some tweaking needed: PySpark recommendation: executor core = 1 - to give DL framework ful access over all resources.
Streaming data through DL:

- cold layer (HDFS, S3…)
- local storage (files..)
- In memory (RDD, DF)

DL commonly used with GPU. GPU is challenging, still. Simplifying: Docker image with GPU SDK & preinstall GPU drivers on instances.
DL&GPU blog on databricks.com

HOW TO INTEGRATE SPARK MLLIB AND APACHE SOLR TO BUILD REAL-TIME ENTITY TYPE RECOGNITION SYSTEM FOR BETTER QUERY UNDERSTANDING

2:40 PM – 3:10 PM
Khalifeh AlJadda from Careerbuilder
link
ETR- Entity Type Recognition -

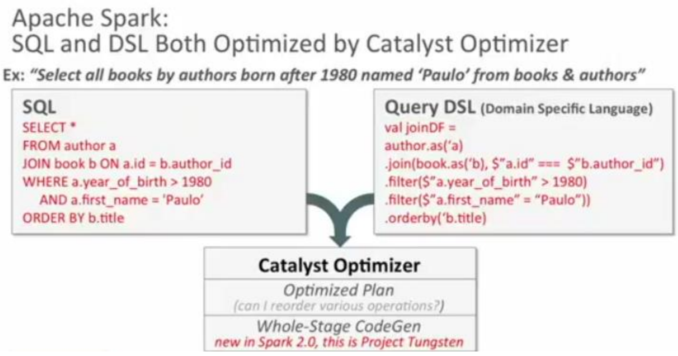
1) Identify regions in text the entities (person, place, name…)
2) Classify the recognized entities

Prior work: Wikipedia data for bag of words and vocabularies
Wikipedia index in SolR (title, length, text, categories)
Word2Vec is used on top of all job postings to generate synonyms or related keywords for each term.
SVM classifier is used to predict the type of entity (skill, job title, location…) - based on Wikipedia and data from Word2Vec.
Research paper on the topic: https://arxiv.org/pdf/1604.00933.pdf
Basing on n-grams.

Sponsored Sessions - ROOM 311 (3/13)

SPARK SQL: ANOTHER 16X FASTER AFTER TUNGSTEN

2:40 PM – 2:55 PM
Brad Carlile from Oracle
link
video
Modern analytics is about your data and external data.
Making JVM faster not Java is the key – also for Spark and Scala.
Sales pitch on SPARC and comparison how it is faster than Spark.

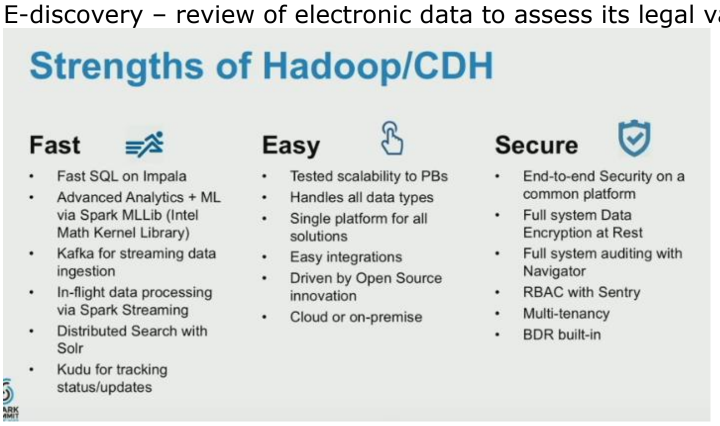


CORNAMI ACCELERATES PERFORMANCE ON SPARK

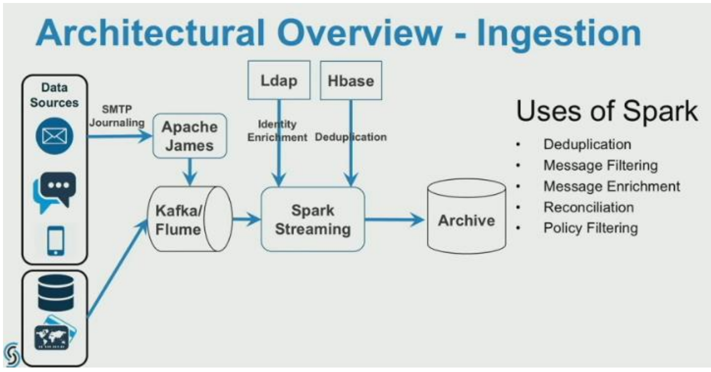
2:55 PM – 3:10 PM

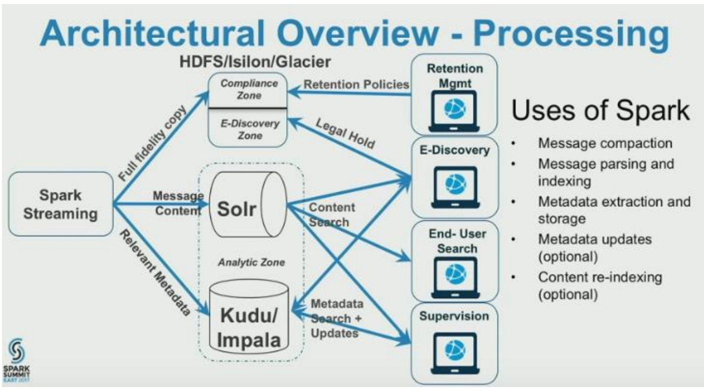
Paul Master from Cornami
link
video
Demo about sea of small cores is faster than Intels solutions with normal cores.
The product offers millions of cores. Scala or Java APIs available.

SOLVING REAL PROBLEMS WITH APACHE SPARK: ARCHIVING, E-DISCOVERY, AND SUPERVISION
11:40 AM – 12:00 PM
Jordan Volz from Cloudera
link
video
Great talk about architectural options.
Long term storage of data – archiving, Passive or active archive – SQL on it or not. How long to archive it…
E-discovery – review of electronic data to assess its legal value.



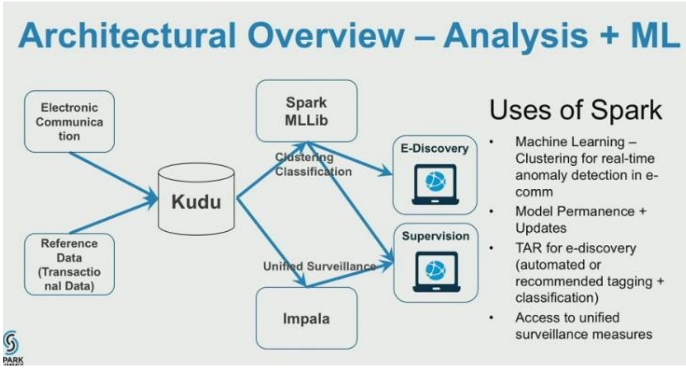
BDR – backup and disaster recovery

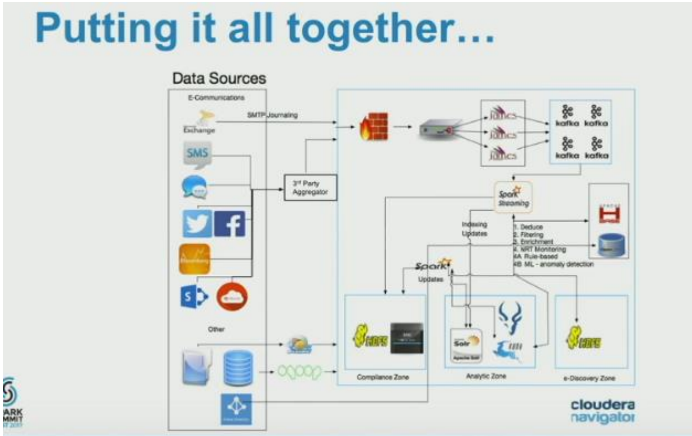




HDFS – original data
Ret Mgmt – rules f. ex how long will some files live

E-Disc Zone – look for legal holes
Solr, Kudu – real time search
Kudu can replace Lambda architecture





Research - ROOM 312 (0/9)

9. Februar (12/50)

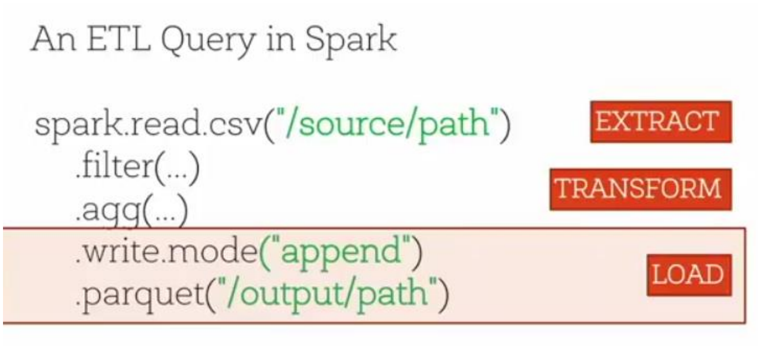
Keynote (0/6)

Spark Ecosystem - Ballroom A (0/9)

Developer - Ballroom B (3/9)

EXCEPTIONS ARE THE NORM: DEALING WITH BAD ACTORS IN ETL

11:00 AM – 11:30 AM
Sameer Agarwal from Databricks
link
video



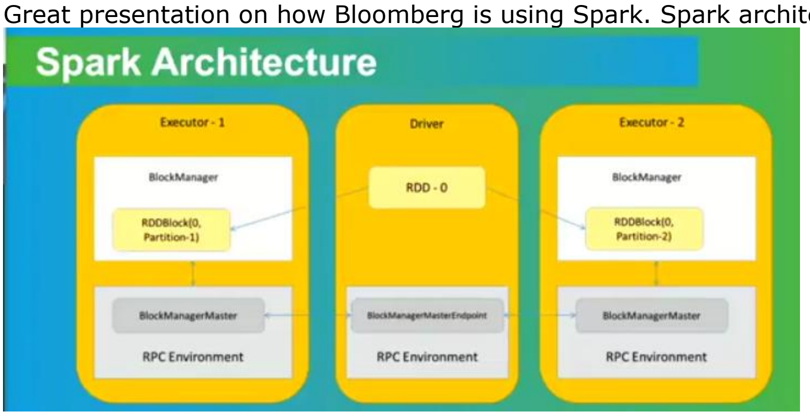
One can skip corrupt files and records. Showing how to handle corrupt records. Very useful and with code examples.
What to expect in Spark 2.2 and 2.3.
Python is the most popular ETL language.

ROBUST AND SCALABLE ETL OVER CLOUD STORAGE WITH SPARK

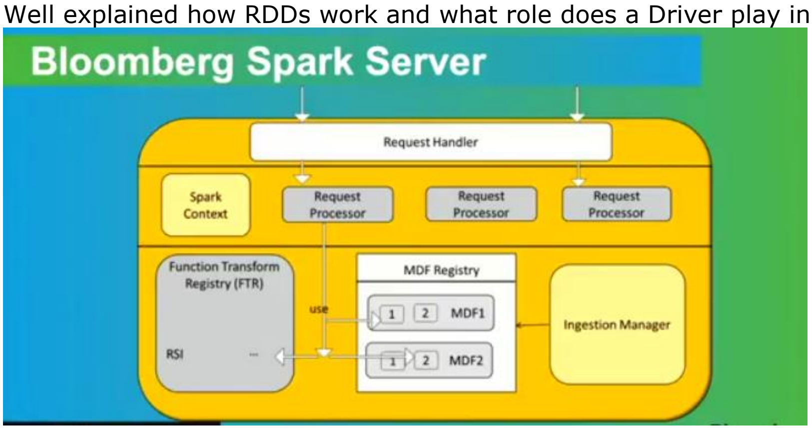
2:00 PM – 2:30 PM
Eric Liang from Databricks
link
video
Why use cloud storage over HDFS – comparing s3 against HDFS.
S3 is 4x cheaper than HDFS on EBS.
S3 is fully managed – means less engineers.
A technical explanation of how to do ETL with Spark on S3, the difficulties one can meet and available solutions on the market to solve these issues.
S3 still deals with some inconsistency, like files not deleted immediately.

SPARK AND ONLINE ANALYTICS

2:40 PM – 3:10 PM
Shubham Chopra from Bloomberg
link
video
Great presentation on how Bloomberg is using Spark. Spark architecture explained.



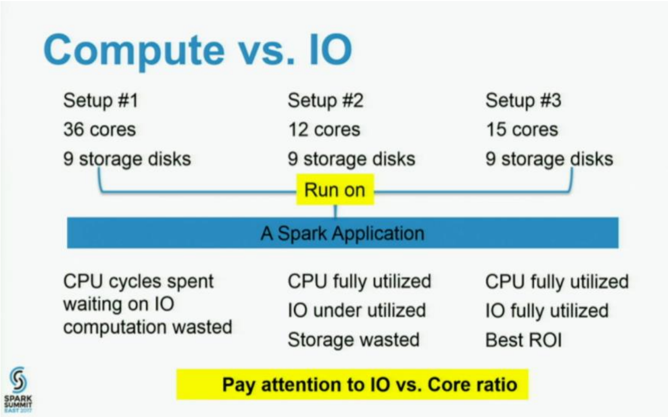
Well explained how RDDs work and what role does a Driver play in the Spark architecture.



Spark experience and Use cases - Ballroom C (1/9)

ACCELERATING SPARK GENOME SEQUENCING IN CLOUD—A DATA DRIVEN APPROACH, CASE STUDIES AND BEYOND

4:20 PM – 4:50 PM

Yingqi (Lucy) Lu from Intel
link


Talk about optimizing computational resources. Improving performance in Java or Scala.

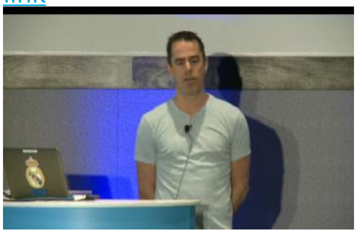
Data Science - ROOM 302/304 (0/9)

Enterprise - ROOM 311 (8/8)

REAL-TIME PLATFORM FOR SECOND LOOK BUSINESS USE CASE USING SPARK AND KAFKA

11:00 AM – 11:30 AM
0:08:10
Ivy Lu from CapitalOne
link
Two systems - for batch processing and real time.
Bach data: used Luigi for job scheduling, Hadoop, Python and MongoDB for mobile push notification. Batch processing sends an email to customer about a potential fraud with a delay.
Real Time pipeline, current phase: Spark, Kafka, Postgres, Cassandra, AWS. Sends an email to user as soon as the card is used - alert sent before the card is put back in the pocket.
Spark is for the majority of logic.
Microservice based - makes it distributed and jobs are decoupled.
Zero data loss challenge (solved with Kafka offset) is described by the speaker. Offsets are sent to Zookeeper - small amounts of data.
They set up Word Cloud of email feedback.
Around 20 people on the team.

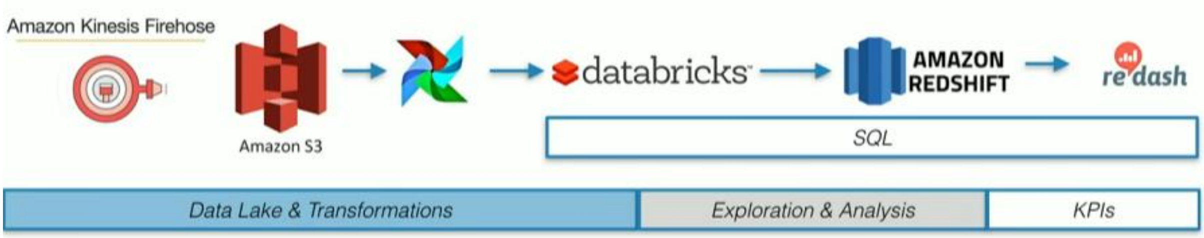
SCALING THROUGH SIMPLICITY—HOW A 300 MILLION USER CHAT APP REDUCED DATA ENGINEERING EFFORTS BY 70%

11:40 AM – 12:10 PM
0:44:00
Joel Cumming from Kik
link

300M users.
Speaker talks about 8 changes they made.
Build a Data Lake in S3. 3 tiers:
1) Raw data
2) Strong type schema, cleaned, reliable data
3) Data for analysis

300 node in EMR -> moved to Databricks, no change in code and it gave faster results.

Collaboration via notebooks.
Used Airflow (open source tool from Airbnb) for orchestration - starting jobs in Spark, etc.
Reporting done in redash (https://redash.io/) - moved from Tableau. Free and open source tool.
With introducing those 8 steps, they saved 70% of the resource time.

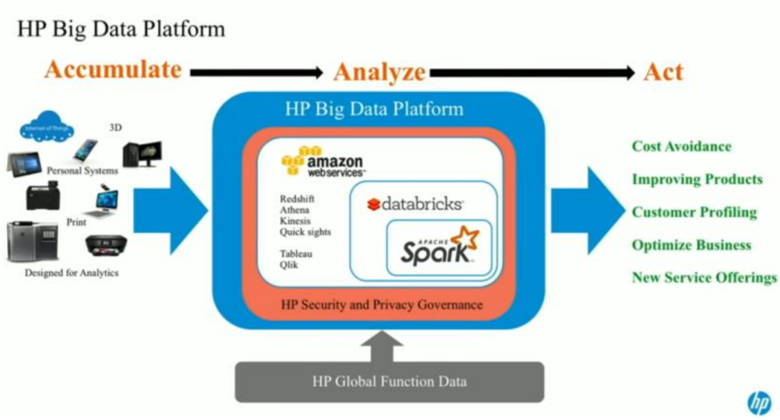
Workflow after the 8 changes were introduced (speaker calls it v.2):



Future research: Spark as a DW? Structured streaming for easier storing of data in DW.

UNLOCKING VALUE IN DEVICE DATA USING SPARK

12:20 PM – 12:50 PM
John Landry from HP Inc.
[link]
[video]



Data from 20M devices moved into S3. Data is brought in in JSON. Stored in Parquet, used first ORC and moved slowly towards Parquet because of optimization. Machine Learning used for cleaning data.
Data Scientists mostly use SQL (SparkSQL).
Spark with Databricks as Big Data platform.

MODELING CATASTROPHIC EVENTS IN SPARK

2:00 PM – 2:30 PM
Georg Hofmann, Shuai Zheng from Validus Research
[link]
Insurance risk analytics.
Using Cat models. These models generate big data, not start them.
MapReduce cluster on EMR. Cluster 5 x r3*8xlarge
They went from MapReduce to Spark. Technical comparison between MapReduce and Spark. Explaining it with numbers and facts. Moving to Spark made it 30%-50% faster, 10-30 times bigger throughput for same cost. They went from r3*8xlarge instances to c3*8xlarge.
Spark gives them:

- high code quality, less code
- more options for architecture design

Quite detailed MapReduce-Spark comparison on AWS.

R&D TO PRODUCT PIPELINE USING APACHE SPARK IN ADTECH

2:40 PM – 3:10 PM
Maximo Gurmendez, Saket Mengle and Sunanda Parthasarathy from Dataxu, Inc
link
Moved from Hadoop to Spark in 2016. Adtech (marketing) company.
A walk through of data science example in Databricks notebook.
Spark made Java/Hadoop easier, less complex because of various languages and working on smaller instance for testing.
Data is stored in S3, Spark loads data, does the job and results are saved in SR buckets for reporting tools to pick them up.
Very practical presentation, explains the lineage of the data, shows the code in 2 notebooks.

FIS: ACCELERATING DIGITAL INTELLIGENCE IN FINTECH

3:20 PM – 3:50 PM
1:27:00
Aaron Colcord from FIS Global
Combination of Lambda architecture with Star schema from BI world and ETL.
Star schema for data on the disk.
Velocity challenge: vertical scaling gave them 5% increase. They went with Spark and horizontal scaling.
They did not go with Hadoop because it is lower level programming. Spark was better because it gives you more focus on high-level programming, data is stored externally, and resources can be used for other jobs.
With Spark, they put together Batch and Speed layer in Lambda architecture.

DISTRIBUTED REAL-TIME STREAM PROCESSING: WHY AND HOW

4:20 PM – 4:50 PM
Petr Zapletal from CakeSolutions
Very detailed and structured talk about streaming solutions on the market. Apache Beam is mentioned as soon-to-be a very good streaming tool.
Practical examples on how to stream with different streaming tools.

HIGH RESOLUTION ENERGY MODELING THAT SCALES WITH APACHE SPARK 2.0

5:00 PM – 5:30 PM
Jonathan Farland from DNV GL
Heavy SparkR user, positive about the package. Combines with Python. Very practical with some R code. Energy sector.