Welcome

GMS Hadoop 1 Day Workshop

A Hands-On Introduction to Hadoop

by Daniel Sikar
dsikar@gmail.com

42 Rue d'Enghien,
75010 Paris

December 2015

Welcome

Technical Training

+

Research and Development

Every can and should contribute

Goals

In our business domain - financial risk calculation area - the workflow usually consists of several steps: some market data retrieval and manipulation, some transaction pricing based on a portfolio and the market data, then some aggregation of the results typically. Pricing step, even if distributed, produces large pricing cubes which may be hard to move around, so maybe it's good to keep them when they were produced. Question for the hadoop based solution would be : is it possible to run specific jobs on specific nodes? For example, we priced a subset of the portfolio on a node, and then we want to run aggregation of this subset on the same node (because the pricing cube should physically reside there). So maybe it would be good to have a two-step workflow with two tasks, where the latter one uses the results produced by the former one.

Youssef Allaoui

More Goals?

Questions we are trying to answer

Questions we are trying to answer

Q: How fast can be HDFS? E.g. who long would it take to move 500 GB of data in Gigabit Ethernet network from one node to another?

Questions we are trying to answer

Q: Can you control how HDFS distributes files? Can you pre-fetch data on a specific node on which you consider to run your job.

Questions we are trying to answer

Q: Is there a way to organize workflows of tasks (simply put a chain, in our case)? Or you have to use Hadoop only as a low-level environment for running jobs?

Questions we are trying to answer

Q: Does Hadoop streaming slow down the execution? How much?

Questions we are trying to answer

Q: Is Microsoft Hadoop SDK mature enough to use in production? Can it be used with standard Hadoop?

Questions we are trying to answer

Q: Does Hadoop monitoring allows estimate current actual loads of nodes?

Questions we are trying to answer

Q: Does Hadoop have a security model? How does it deal with authentication, authorization etc.? Which protocols and standards are used for that?

Questions we are trying to answer

Q: In case of non-java applications what is the model of delivering the running code to the actual location of execution?

Questions we are trying to answer

Q: Does Hadoop allows running long-running tasks that can last very long persisting some kind of context and serving incoming requests?