Hadoop User Group Italy Meeting
Incontro con Marcel Kornacker, Impala Tech lead

19 maggio 2015 dalle 17:45 alle 19:30

Università degli Studi di Milano, Dipartimento di Informatica
Via Comelico, 39 - 20135 Milano (Sala Lauree)

PARTECIPAZIONE GRATUITA. REGISTRAZIONE OBBLIGATORIA.
http://bit.ly/HUG-Italy-meetup

Un'occasione unica sia per **NEWCOMERS** per conoscere le principali soluzioni open source per i **Big Data** sia per **ESPERTI** per incontrare uno dei pricipali Software Engineer di **Hadoop**.

Cloudera Impala è il query engine real-time open source progettato per sfruttare la flessibilità e scalabilità di **Apache Hadoop** e consentire interrogazioni ad elaborazione parallela e massiva ad elevate prestazioni.

Friction-free ETL: Automating data transformation with Impala
Marcel Kornacker - Architect & Tech Lead di Impala, Cloudera Inc.

As data is ingested into Apache Hadoop at an increasing rate from a diverse range of data sources, it is becoming more and more important for users that new data be accessible for analysis as quickly as possible - because "data freshness" can have a direct impact on business results.

In the traditional ETL process, raw data is transformed from the source into a target schema, possibly requiring flattening and condensing, and then loaded into an MPP DBMS. However, this approach has multiple drawbacks that make it unsuitable for real-time, "at-source" analytics - for example, the "ETL lag" reduces data freshness, and the inherent complexity of the process makes it costly to deploy and maintain, and reduces the speed at which new analytic applications can be introduced.

In this talk, attendees will learn about Impala's approach to on-the-fly, automatic data transformation, which in conjunction with the ability to handle nested structures such as JSON and XML documents, addresses the needs of at-source analytics - including direct querying of your input schema, immediate querying of data as it lands in HDFS, and high performance on par with specialized engines. This performance level is attained in spite of the most challenging and diverse input formats, which are addressed through an automated background conversion process into Parquet, the high-performance, open source columnar format that has been widely adopted across the Hadoop ecosystem.

L'intervento si terrà in lingua inglese.

Con il supporto di: