GraphBuilder: A Scalable Graph ETL Framework

Authors: Nilesh Jain, Guangdeng Liao, Theodore Willke
Presented By: Kushal Datta
Legal Notices

- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL’S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY WHATSOEVER, AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL® PRODUCTS INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT. INTEL PRODUCTS ARE NOT INTENDED FOR USE IN MEDICAL, LIFE SAVING, OR LIFE SUSTAINING APPLICATIONS.

- Intel may make changes to specifications and product descriptions at any time, without notice.

- All products, dates, and figures specified are preliminary based on current expectations, and are subject to change without notice.

- Intel, processors, chipsets, and desktop boards may contain design defects or errors known as errata, which may cause the product to deviate from published specifications. Current characterized errata are available on request.

- Code names featured are used internally within Intel to identify products that are in development and not yet publicly announced for release. Customers, licensees and other third parties are not authorized by Intel to use code names in advertising, promotion or marketing of any product or services and any such use of Intel’s internal code names is at the sole risk of the user.

- Performance tests and ratings are measured using specific computer systems and/or components and reflect the approximate performance of Intel products as measured by those tests. Any difference in system hardware or software design or configuration may affect actual performance.

- Intel, Intel Inside, and the Intel logo are trademarks of Intel Corporation in the United States and other countries.

- *Other names and brands may be claimed as the property of others.

- Copyright © 2013 Intel Corporation.
Graphs

Graph is an abstract representation of a set of objects (vertices) and the relationships between them (edges) which can be either directed or undirected.
Graphs are omnipresent...

Big in size and rich in metadata

Intel Labs/Graph Analytics Operation
Graphs are omnipresent...

Big in size and rich in metadata

Intel Labs/Graph Analytics Operation
Graphs are omnipresent...

- Social Networks: 1B Users, 140B Friendships
- Online Services: 50K Users, 17K Movies, 1B Ratings
- Science: Large Biological Cell Networks
- E-commerce: Millions of Products & Users
- Human Brain: 100B Neurons, 100T Connections

Big in size and rich in metadata

Intel Labs/Graph Analytics Operation

Functional connections in the brain obtained from correlating activation patterns in fMRI
(http://www.gipsa-lab.grenoble-inp.fr/~sophie.achard/recherches_en.html)
Graphs are Essential to Data Mining and Machine Learning

• Identify influential people and information
• Find communities
• Understand people’s shared interests
• Model complex data dependencies
“I spend more than half of my time integrating, cleansing and transforming data without doing any actual analysis. Most of the time I’m lucky if I get to do any analysis at all.”

Anonymous Data Scientist from Jeff Heer’s (Stanford) interview study, 2012
Graph Construction

Social Networking Data

Feature Extraction

Graph Construction Data Parallel

Graph Formation

Graph

Hadoop is perfect for graph construction!

Intel Labs/Graph Analytics Operation
GraphBuilder makes it easy.

- Fills a hole in the ecosystem
- Written in Java for easy use in Hadoop MapReduce and applications
- Offloads domain expertise

GraphBuilder Library

Raw Data

Parsing

Tokenization or Feature Extraction

Edge List Generation

E, V Data Tabulation

Graph Normalization

Graph Checks & Transforms

Graph Partitioning & Serialization

To Graph Processing

App-Specific Code

Intel Labs/Graph Analytics Operation
GraphBuilder Stack

Graph Partitioning
Data Normalization
Built-in Parsers
Tabulations
Framework Adapter
User Defined
Graph Transformation
GraphBuilder
MapReduce
Distributed Graph Computation (GLv2)
HDFS

GraphBuilder app
ML App

User defined parser
User defined functions

Intel Labs/Graph Analytics Operation
GraphBuilder Scales Linearly

Overall Time (Seconds)

Dataset Size

- Finalization
- Partitioning
- Pre-processing/Graph Formation

Intel
Graph Builder source code is available at www.01.org/graphbuilder

Contact: kushal.datta@intel.com

Thanks!