SIGMOD GRADES 2013

GraphBuilder: A Scalable Graph ETL Framework

Large Scale Graph Construction using Apache™ Hadoop™

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

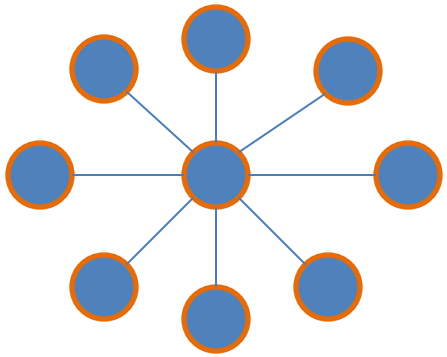
Intel Labs/Graph Analytics Operation

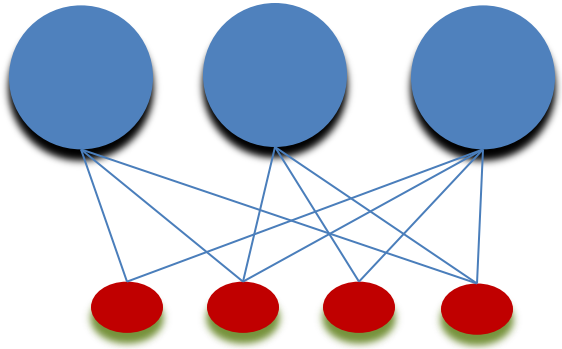
Legal Notices

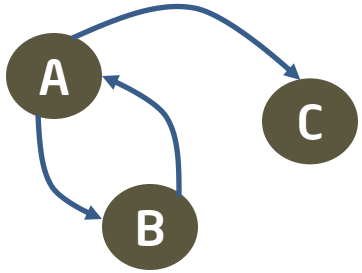
- INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL® PRODUCTS. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS, INTEL ASSUMES NO LIABILITY 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.
-

Graphs

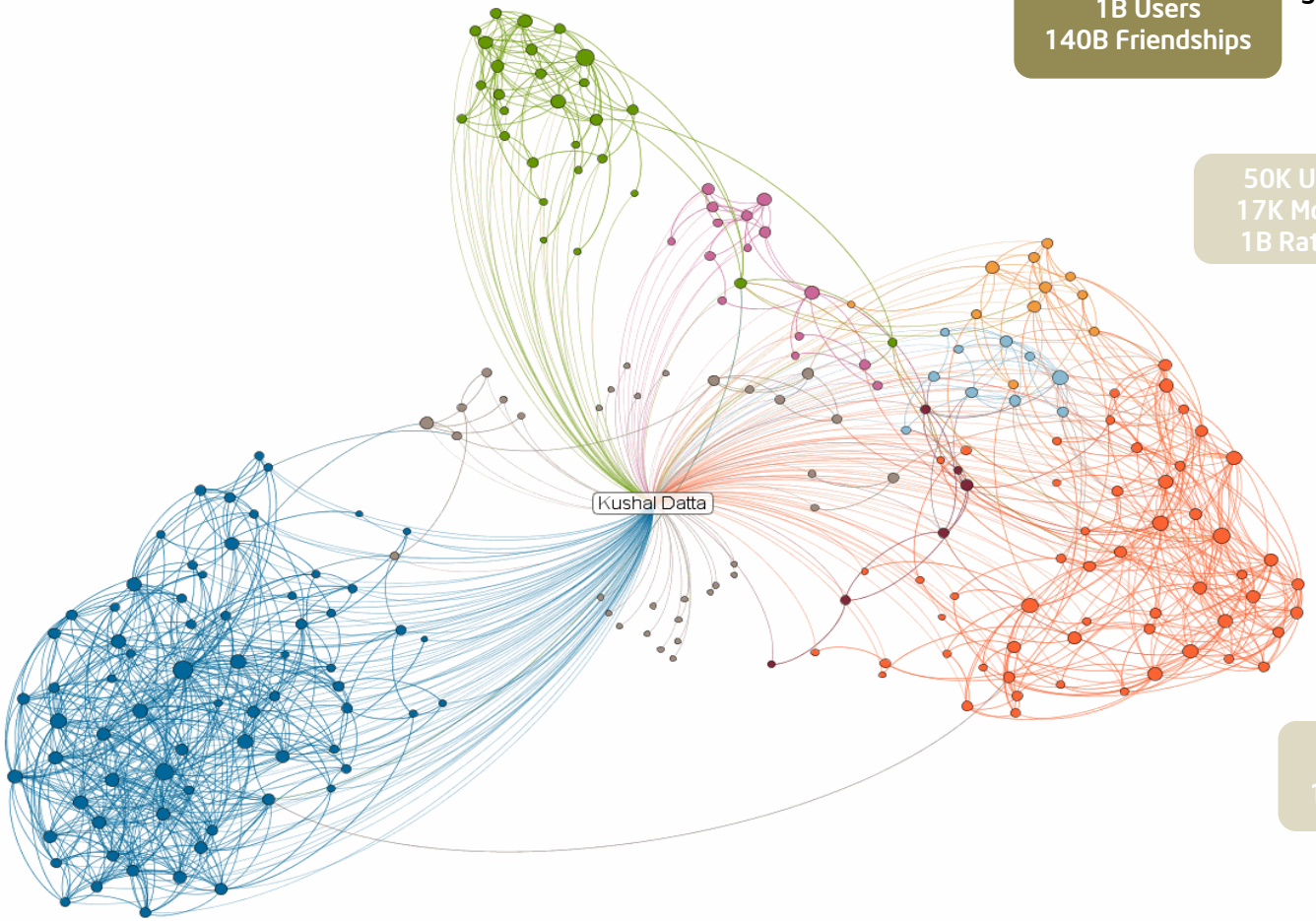
Graph is an abstract representation of a set of objects (vertices) and the relationships between them (edges) which can be either directed or undirected







Graphs are omnipresent…



1B Users
140B Friendships
Social Networks

50K Users
17K Movies
1B Ratings
Online Services

Large Biological
Cell Networks
Science

Millions of
Products & Users
E-commerce

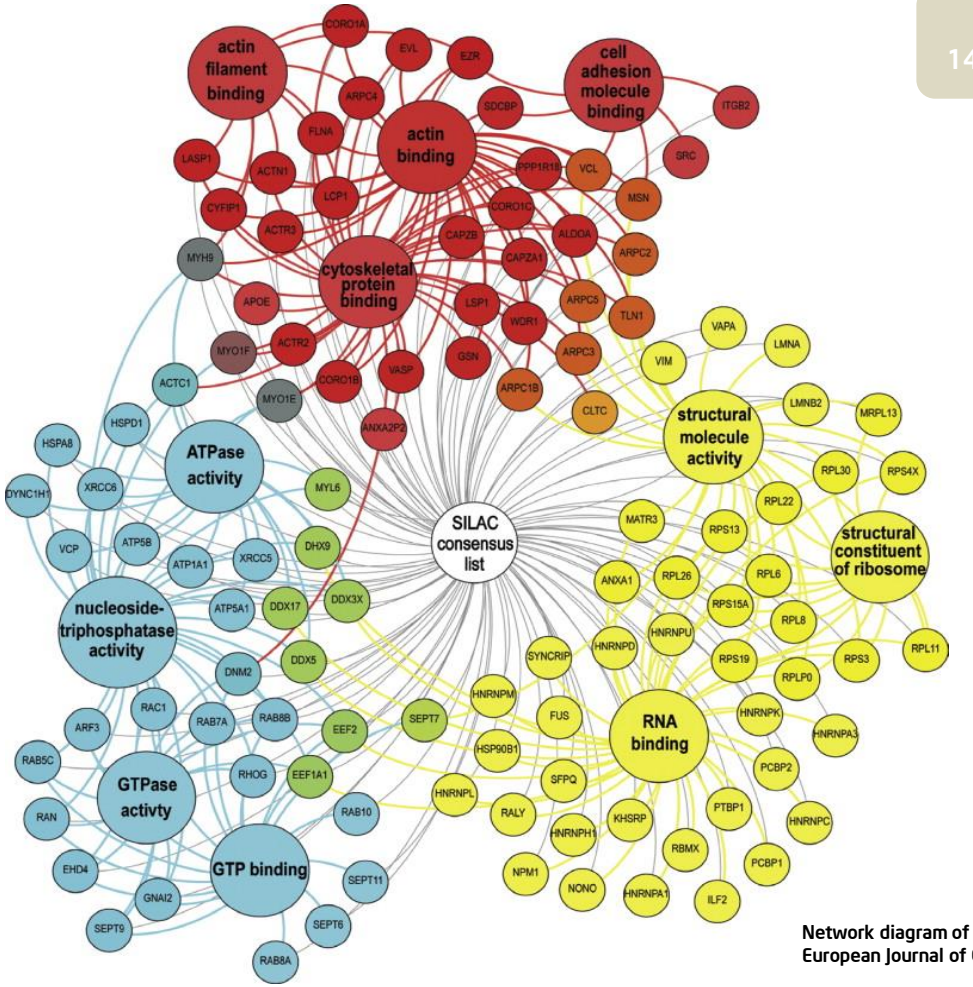
100B Neurons
100T Connections
Human Brain

Big in size and rich in metadata

Intel Labs/Graph Analytics Operation

intel

Graphs are omnipresent…



Network diagram of consensus list proteins
European Journal of Cell Biology, Vol 91, Dec 2012

1B Users
140B Friendships
Social Networks

50K Users
17K Movies
1B Ratings
Online Services

Large Biological Cell Networks
Science

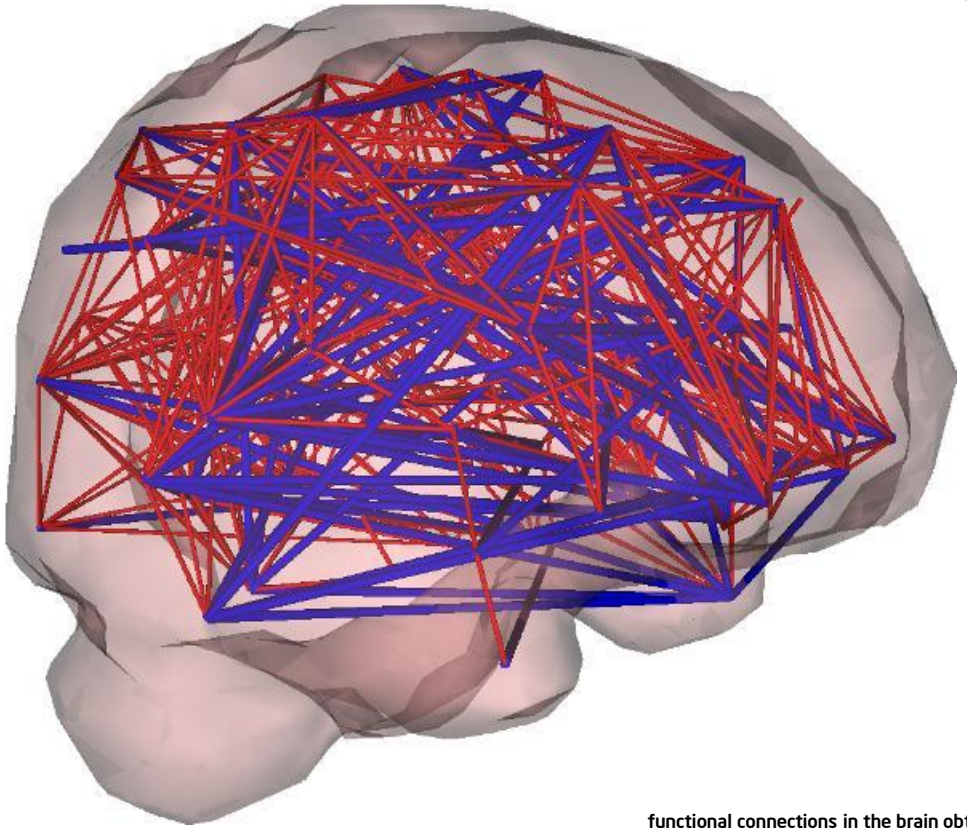
Millions of Products & Users
E-commerce

100B Neurons
100T Connections
Human Brain

Big in size and rich in metadata

Intel Labs/Graph Analytics Operation

Graphs are omnipresent…



1B Users
140B Friendships — Social Networks

50K Users
17K Movies
1B Ratings — Online Services

Large Biological
Cell Networks — Science

Millions of
Products & Users — E-commerce

100B Neurons
100T Connections — **Human Brain**

functional connections in the brain obtained from correlating activation patterns in fMRI
(http://www.gipsa-lab.grenoble-inp.fr/~sophie.achard/recherches_en.html

Big in size and rich in metadata

Intel Labs/Graph Analytics Operation

intel

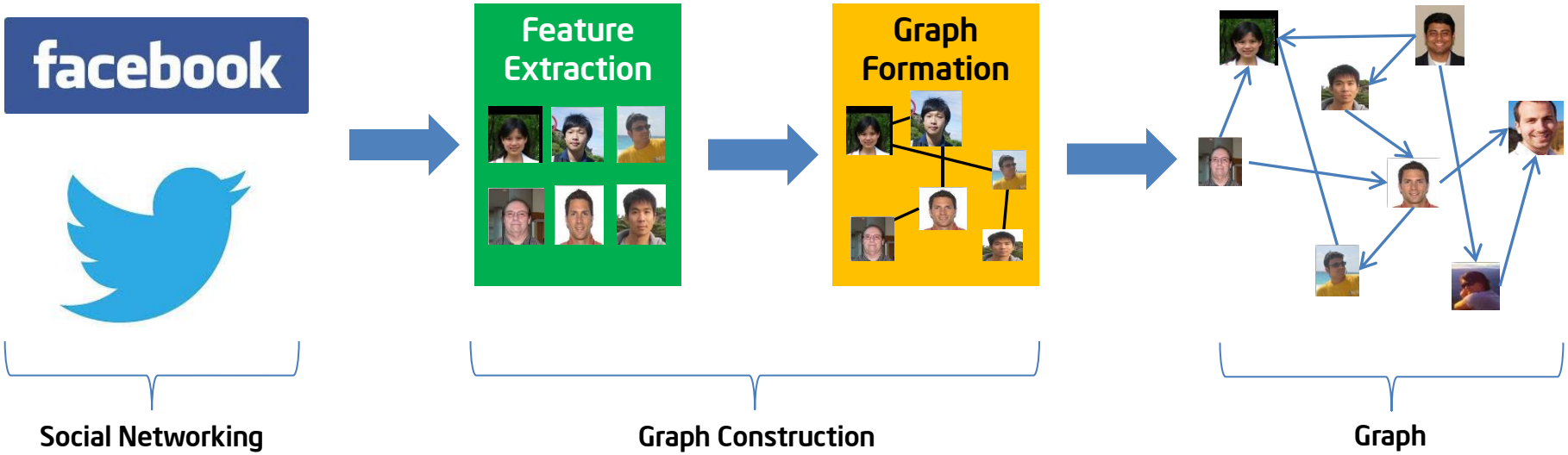
Graphs are Essential to Data Mining and Machine Learning

- 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

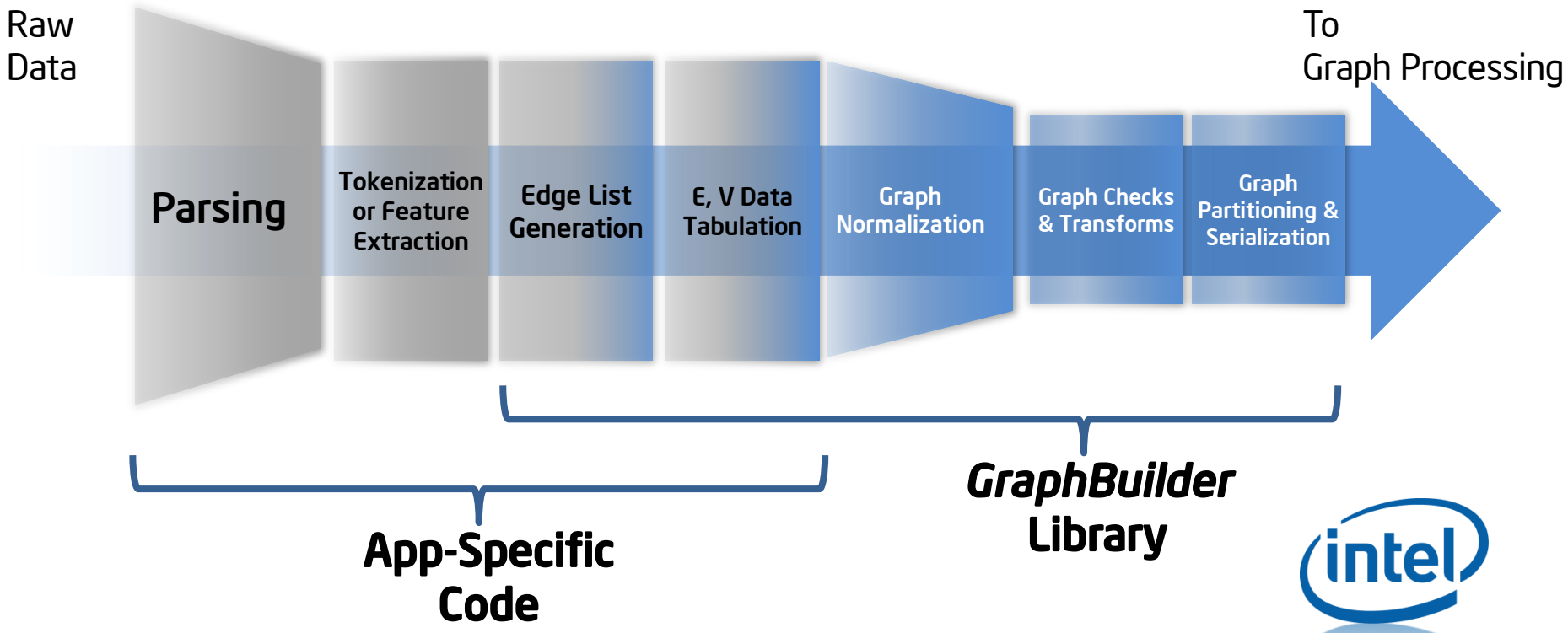
Graph Construction



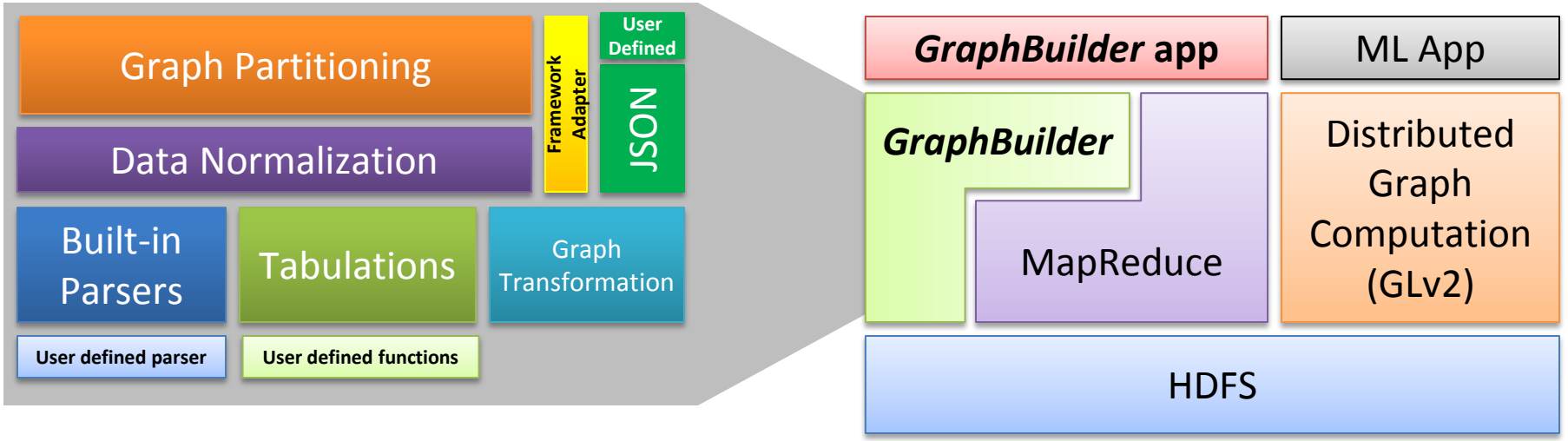
Hadoop is perfect for graph construction!

GraphBuilder makes it easy.

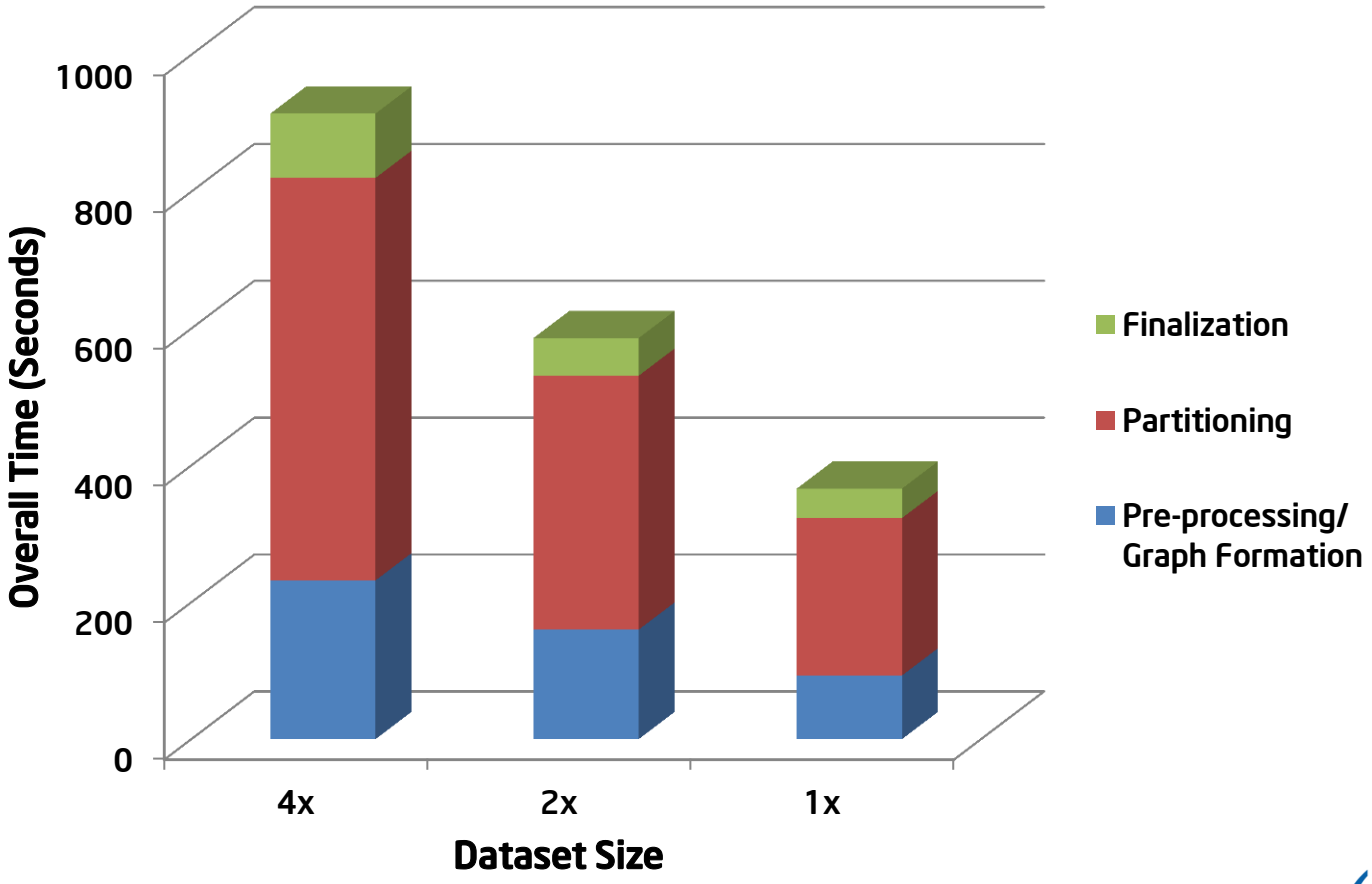
- Fills a hole in the ecosystem
- Written in Java for easy use in Hadoop MapReduce and applications
- Offloads domain expertise



GraphBuilder Stack



GraphBuilder Scales Linearly



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

Contact: kushal.datta@intel.com

Thanks!