An experimental comparison of graph processing systems

Payas Rajan • UC - Riverside • 10.27.2016
Overview

Why graphs?

Frameworks

- Apache Giraph
- Apache GraphX
- Pregelix

Comparison parameters
Why graphs?

1. They are the ‘natural’ representation.
2. They allow powerful operations.

- Web graph - PageRank
- Road networks - Routing
- Social networks - Facebook connections
Why graphs?

1. They are the ‘natural’ representation.
2. They allow powerful operations.

- Web graph - PageRank
- Road networks - Routing
- Social networks - Facebook connections
Why graphs?

1. They are the ‘natural’ representation.
2. They allow powerful operations.

- Web graph - PageRank
- Road networks - Routing
- Social networks - Facebook connections
How big are graphs?

- Web: 8.53 Billion pages in 2012
- Facebook active users: 1.01 Billion
- de Bruijn graph: 3 Billion nodes
Some unconventional uses...

Microsoft Graph:
The easiest way to call Microsoft APIs

Power your app with rich, intelligent data from Office 365 and Azure AD — all from a single endpoint.

See overview  Try the API

Source: https://graph.microsoft.io/en-us/
Some unconventional uses...

Source: https://developers.facebook.com/docs/graph-api
Some unconventional uses...

Source: https://developer.musicgraph.com/
Some unconventional uses...

Source: https://turi.com/
Overview

Why graphs?

Frameworks

- Apache Giraph
- Apache GraphX
- Pregelix

Comparison parameters
Apache Giraph

- Iterative graph processing framework, built on top of **Apache Hadoop**.
- Uses the Bulk-Synchronous-Parallel approach.
- Input is a graph composed of vertices and directed edges.
- “Think like a vertex.”

Source: http://giraph.apache.org/intro.html
GraphX

- Extends the Spark RDD by introducing a new Graph abstraction: a directed multigraph with properties attached to each vertex and edge.

Source: http://spark.apache.org/graphx/
Pregelix

- Uses a set-oriented, dataflow approach to implement BSP model.
- Treat messages and vertex states like tuples in a schema.
- Message exchanges now are the joins of schemas.
- Use existing distributed dataflow execution engines: Hyracks

Source: author’s slides
Overview

Why graphs?

Frameworks

- Apache Giraph
- Apache GraphX
- Pregelix

Comparison parameters
Why the comparison?

End-to-end PageRank performance (20 iterations, 3.7B edges)

Source: GraphX website

Source: Pregelix: Big(ger) Graph Analytics on A Dataflow Engine (Bu et al. VLDB 2015)
Why the comparison?

http://spark.apache.org/faq.html

Does my data need to fit in memory to use Spark?

No. Spark's operators spill data to disk if it does not fit in memory, allowing it to run well on any sized data. Likewise, cached datasets that do not fit in memory are either spilled to disk or recomputed on the fly when needed, as determined by the RDD's storage level.

End-to-end PageRank performance (20 iterations, 3.7B edges)

Source: GraphX website

Source: Pregelix: Big(ger) Graph Analytics on A Dataflow Engine (Bu et al. VLDB 2015)
Comparison parameters

- **Datasets:**
  - Road network - spatialHadoop (717 M records, 137GB)
  - Wikipedia

- **Algorithms:**
  - PageRank
  - Finding triangles
  - Strongly Connected Components

- **Cluster:**
  - Two machines, Intel Xeon, 64GB RAM
Thank You :)