Vinay Santhosham

ID: 16264245

**Pregel**

**Abstract:**

* Computing problems on graphs
* Each graph has billions of vertices and edges
* Sequence of iterations
* Send messages to vertices via edges
* Modify their own state
* Vertex centric approach

**Introduction**

* Web graph a example for analysis and research
* The webgraph is a directed graph, whose vertices correspond to the pages of the

WWW

* Directed graph : Edges -> vertices
* Past Research : routes, paths of disease outbreak, published citation - > graph analysis
* Shortest path, difference in page rank
* Processing large graphs is challenging
* Poor memory access
* Little work per vertex
* Locality issue

**Features**

* Scalable
* Fault-tolerant
* Flexible API

**C++ API**

* Pregel program -> sub classing predefined Vertex class.
* Compute() is vertex associated value by GetValue() or modify by MutableValue()
* Values of edges can be inspected and modified.

**C++ API- Message Passing**

* Each message consists of a value and name of the destination vertex.
* Type of value -> template parameter of Vertex class.

Any number of messages can be sent in a superstep.

* Combiners are used to send messages to another vertex.
* Combiners can be used only for associative and commutative operation.
* Aggregators -> Global communication, monitoring and data. Each vertex can produce a value in a superstep s for the Aggregator to use.

**Topology Mutation**

* Compute() -> modify a graph topology i.e by sending messages, requesting to add or delete vertices or edges.
* Multiple requests to add or delete same vertex oredge in same superstep cause conflicts.
* 2 mechanisms to overcome them:

•Partial ordering

•User defined handlers.

* Partial ordering of mutations can yield deterministic results

• removals taking place before additions

• removal of edges before vertices

•addition of vertices before edges.

* Pregel is designed for Google cluster architecture where each cluster has huge number of PC arranged in racks.
* Clusters are distributed geographically but are interconnected.
* Pregel execute on cluster management where the jobs are scheduled to optimize resource allocation
* Pregel library divides a graph into partitions based on vertex ID
* The default function is hash mod N, where N is the number of partitions.