
Ab initio Session 14
Introduction to Ab Initio
Capgemini
Ab Initio Training
1

- **Concepts of Parallelism**
- **Explanation of Data partitioning**
- **Concept of Repartitioning**

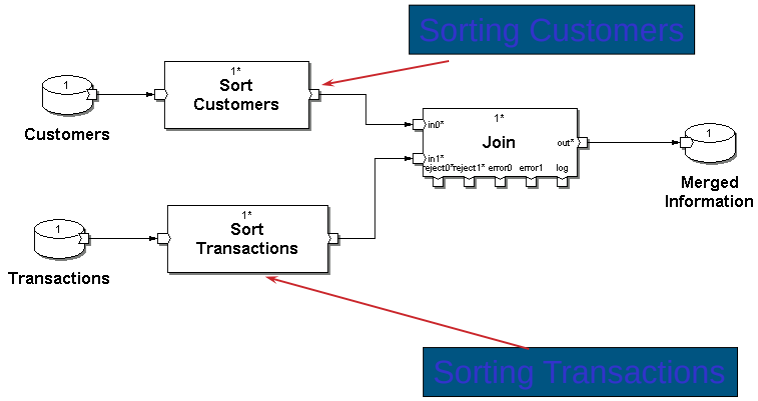
Forms of Parallelism

- **Component parallelism**
- **Pipeline parallelism**
- **Data parallelism**

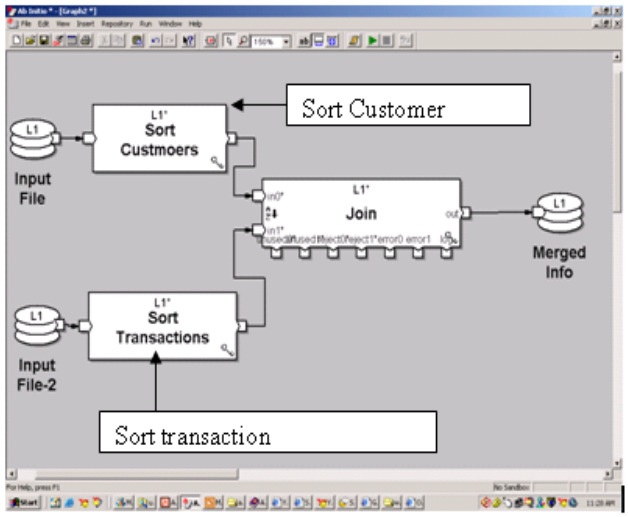
Component parallelism

- A graph with multiple processes running simultaneously on separate data uses component parallelism
- In this two or more components process the records in parallel.

Component Parallelism



CapGemini Ab Initio Training 5


Sort Customer
L1
Sort
Custmoers
Input
File
L1
Join
L1
Merged
Info
L1
Sort
Transactions
Input
File-2
Sort transaction

Pipeline Parallelism

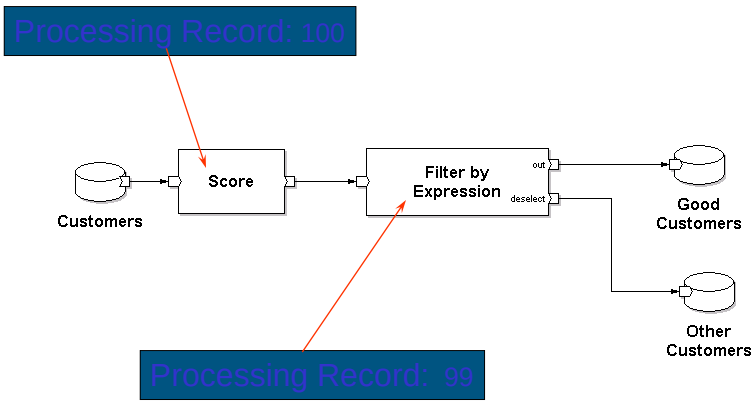
- A graph with multiple components running simultaneously on the same data uses pipeline parallelism. Each component in the pipeline continuously reads from upstream components, processes data, and writes to downstream components. Since a downstream component can process records previously written by an upstream component, both components can operate in parallel. NOTE: To limit the number of components running simultaneously,set phases in the graph.

Pipeline Parallelism-cont.

- Each component in the pipeline continuously reads from upstream components, processes data, and writes to downstream components. Since a downstream component can process records previously written by an upstream component, both components can operate in parallel.

  NOTE: To limit the number of components running simultaneously, set phases in the graph.

Pipeline Parallelism



CapGemini Ab Initio Training 9

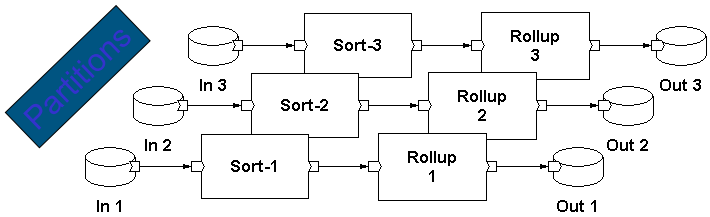
Pipeline Parallelism-cont.

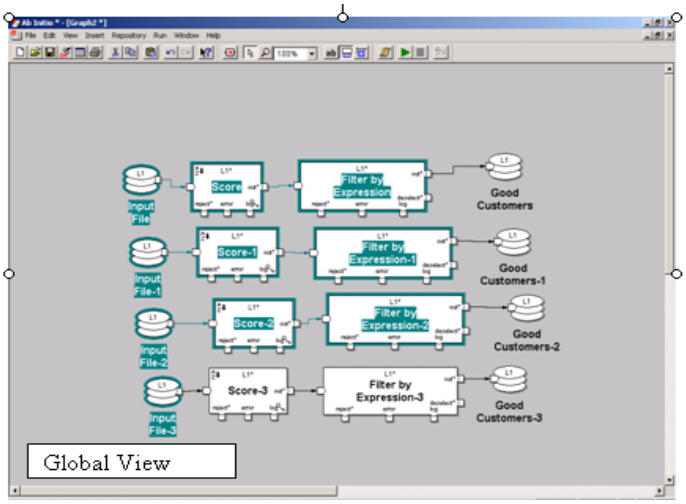
- In this the records are processed in pipeline, i.e. the components do not have to wait for all the records to be processed. The records that got processed are passed to next component in pipeline.

Data Parallelism

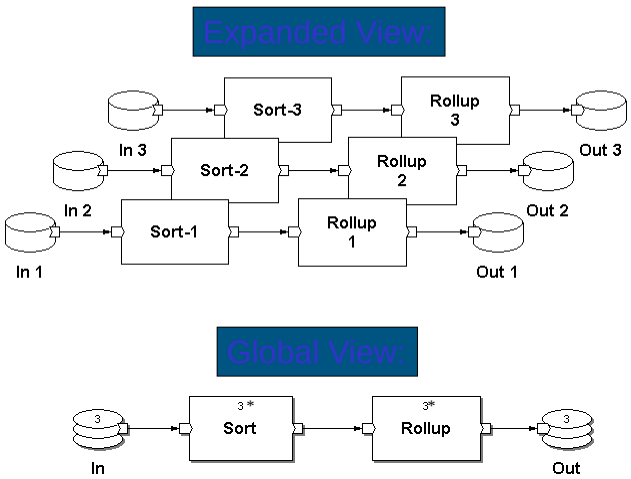
- A graph that deals with data divided into segments and operates on each segment simultaneously uses data parallelism. Nearly all commercial data processing tasks can use data parallelism. To support this form of parallelism, Ab Initio provides Partition components to segment data, and Departition components to merge segmented data back together.
- Partitioning is an example of data parallelism.

Data Parallelism




Input File
Score
Filter by Expression
Good Customers
Input File-1
Score-1
Filter by Expression-1
Good Customers-1
Input File-2
Score-2
Filter by Expression-2
Good Customers-2
Input File-3
Score-3
Filter by Expression-3
Good Customers-3
Global View

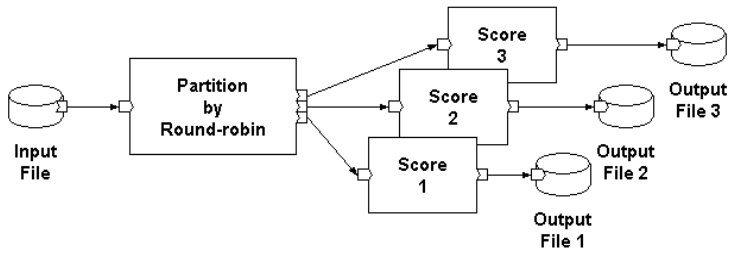
Two Ways of Looking at Data Parallelism



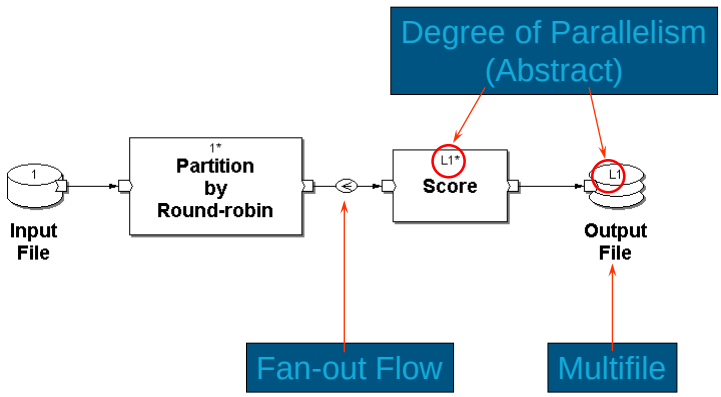
Data Parallelism

- Scales with data.
- Requires *data partitioning*.
- Dependent upon the application, different partitioning methods are available.

A Data Parallel Application: The Expanded View



A Data Parallel Application: The Global View



Thank You