Universidad de los Andes

Ingeniería de Sistemas y Computación

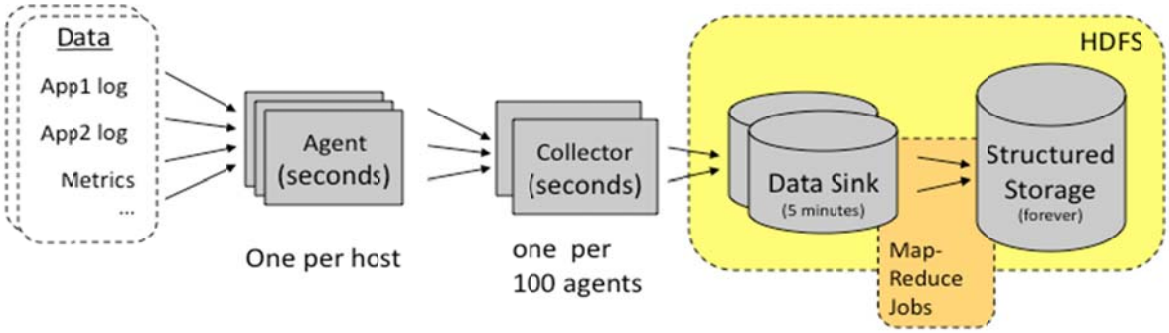
ISIS2503 Arquitectura y diseño de software

Actividad en clase

Instrucciones

1. Leer la descripción de la arquitectura Suro la cual se utiliza en Netflix
2. Responder las siguientes preguntas:
   - A qué atributos de calidad apunta. Explique por qué
   - Cuál de las tácticas de desempeño se aplican en la arquitectura propuesta (ver anexo). Explique por qué

Suro architecture



This architecture aims to provide a flexible and powerful platform for distributed data collection and rapid data processing. The architecture is structured as a pipeline of collection and processing stages. Its principal elements are: agents, collectors and a Hadoop Distributed File System (HDFS).

Agents

- Collect data (data is produced by a command or operations over a file)
- Emit data in *Chunks* in a pre-established frequency. A Chunk is a sequence of bytes, with some metadata
  - Sequence ID is part of the metadata
    - if an agent emits a chunk containing the first 100 bytes from a file, the sequenceID of that Chunk will be 100
    - sequence ID is a parameter so that agents can resume correctly after a crash, and not send redundant data

Collectors

Rather than have each adaptor writing directly to HDFS, data is sent across the network to a *collector* process that does the HDFS writes. Each collector receives data from up to several hundred hosts, and writes all this data to a single *sink file*, which is a Hadoop sequence file of serialized Chunks. Periodically, collectors close their sink files, rename them to mark them available for processing, and resume writing a new file. Data is sent to collectors over HTTP.

Map reduce jobs

Collectors write data in sequence files. This is convenient for rapidly getting data committed to stable storage. But it's less convenient for analysis or finding particular data items. As a result, Suro has a toolbox of MapReduce jobs for organizing and processing incoming data.

Anexo