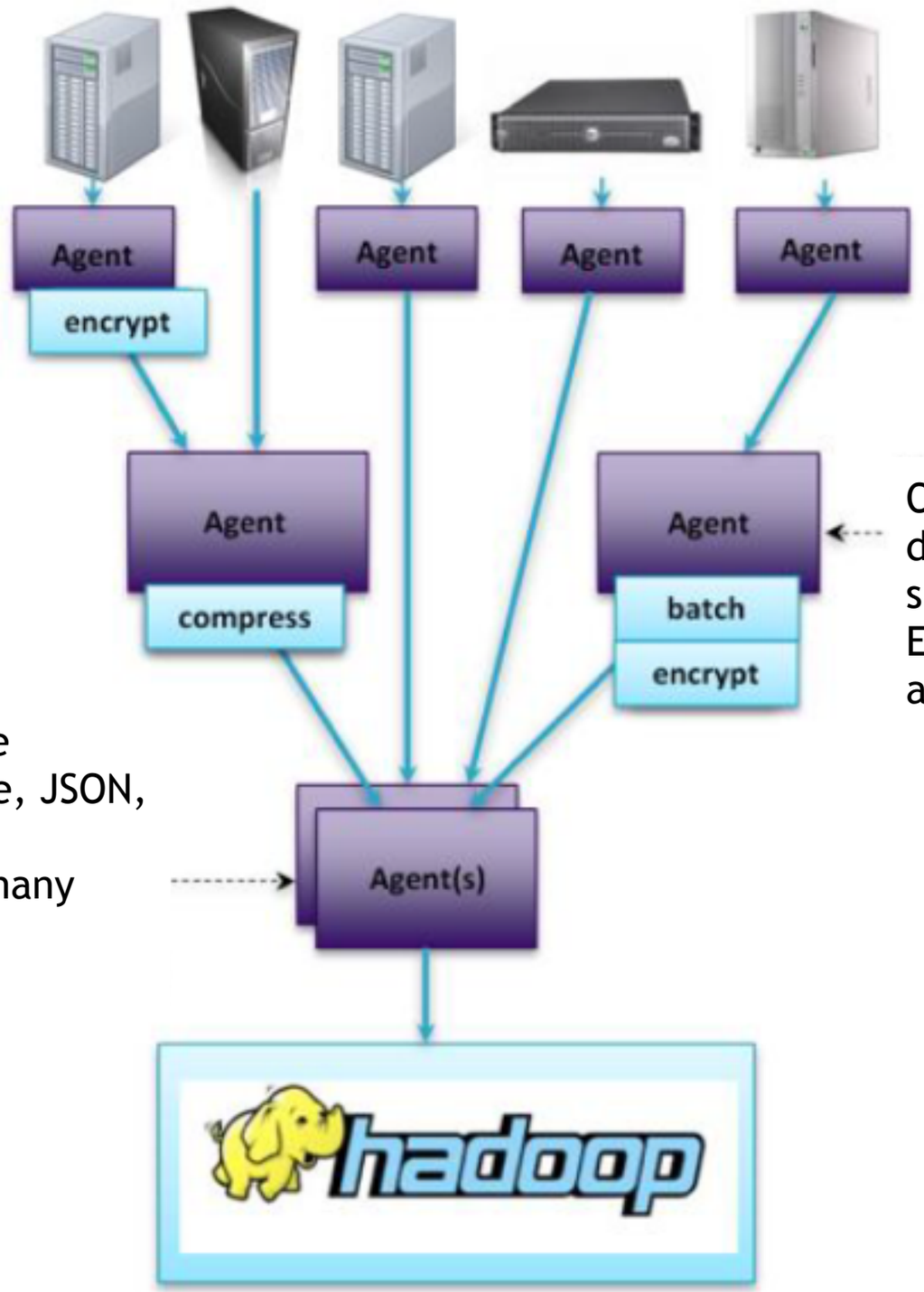
# Flume: Basics

- **Flume is a distributed, reliable, available service for efficiently moving large amounts of data as it is produced**
  - **Ideally suited to gathering logs from multiple systems and inserting them into HDFS as they are generated**

# Flume: Usage Patterns

- **Flume is typically used to ingest log files from real-time systems such as Web servers, firewalls, and mailservers into HDFS**
- **Used in many large organizations, ingesting millions of events per day**

# Flume: High-Level Overview




Optionally pre-process incoming data perform transformations, suppressions, metadata enrichment
Each agent can be configured with an in-memory or durable channel.

Writes to multiple HDFS file formats (text, SequenceFile, JSON, Avro, others)
Parallelized writes across many collectors – as much write throughput as required

# Flume Agent Characteristics

- Each Flume agent has a source and a sink
- Source
  - Tells the node where to receive data from
- Sink
  - Tells the node where to send data to
- Channel
  - A queue between the Source and Sink
  - Can be in memory or 'Durable'

# Flume's Reliability

- Channels provide Flume's reliability
- Memory Channel
  - Data will be lost if power is lost
- Disk-based Channel
  - Disk-base queue guarantees durability of data in face of a power loss
- Data transfer between Agents and Channels is transactional
- Can configure multiple Agents with the same task
  - e.g., 2 Agents doing the job of 1 'collector' - if one agent fails then upstream agents would fail over

# Flume's Scalability

- Scalability
  - The ability to increase system performance linearly - or better - by adding more resources to the system
  - Flume scales horizontally
    - As load increases, more machines can be added to the configuration

# Flume's Extensibility

- Flume can be extended by adding Sources and Sinks to existing storage layers or data platforms
  - General Sources include data from files, syslog, and standard output from any Linux process
  - General Sinks include files on the local filesystem or HDFS
  - You can write your own Sources or Sinks