Atrax

# Introduction

This document discusses the two principle parts of the Atrax web crawler: the Frontier and the Fetcher.

# Definitions

Command & Control (C&C): A designated machine that is used to communicate with all other machines in the cloud. The C&C instance stores information about each crawl job that other instances need in order to operate. Crawl Job specific files are stored in directories with the crawl job name.

# Deployment

These are the steps for deploying Atrax and initializing a crawl job for the first time.

1. Run the crawl job initialization script to setup resources needed by the crawl job.
2. Upload the environment setup scripts for the Frontier and Fetcher to C&C using the Atrax Keeper web service.
3. Upload the Frontier and Fetcher packages to C&C using the Atrax Keeper web service.
4. Create the frontier machine.
5. Make sure that the proxy is started.
6. Upload the crawl job configuration to C&C using the Atrax Keeper web service
7. Start the frontier machine.
8. Request the spot instances.

# Crawler Job Startup

Each time the Frontier starts up it performs the following steps.

1. Download and execute the startup script if C&C has a newer version.
2. Download and execute the environment setup script if C&C has a newer version. Store the modified time of the file on the instance.
3. Download the crawl job configuration from C&C if it is newer. Store the modified time of the file on the instance.
4. Download and execute the packages for the instance type if C&C has a newer version. Store the modified time of the file on the instance.

# Pausing a Crawl Job

Crawl jobs should not be paused for more than 10 days. If the crawl job needs to be stopped for more than 10 days then the stop crawl job procedure should be followed. These steps are performed to pause a crawl job.

1. Terminate all spot instances.
2. Stop the frontier instance.

# Stopping a Crawl Job

1. Terminate all spot instances
2. Save all frontier queues to persistent storage