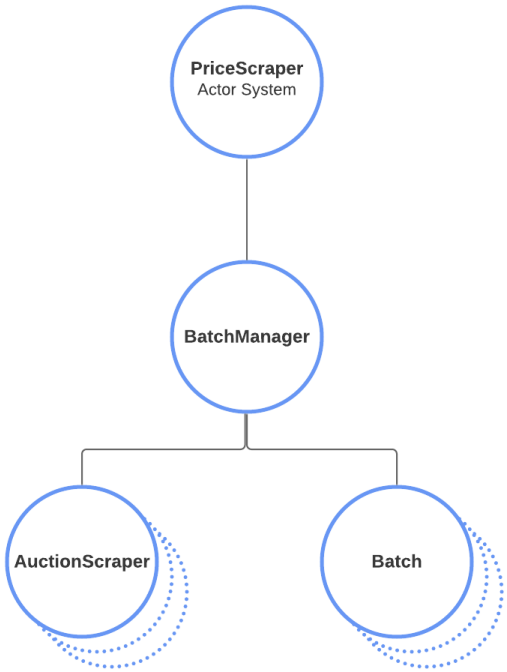
PriceScraper

Table of Contents

1. PriceScraper

The **PriceScraper** is a process that allows to extract auctions along with their characteristics, based on a list of **Batch Specifications**. The price scraper doesn't create isolated auctions, instead it creates what we call a **Batch** that contains a list of auctions, this list of auctions are those auctions that are found on a web page at the time of scraping. Such a web page is for example a page containing the list of stamps for an url provided in a Batch Specification.

We have decided to use Scala and Akka with its system of actors for the implementation, below is an overview of the actor system at the time of writing :



The different components of this actor system are :

PriceScraper: it's the entry point of the application and it creates the actor system and the BatchManager

BatchManager: it communicates with the AuctionScraper to provide the BatchSpecifications and it creates the Batch actors upon reception of messages from the AuctionScraper

AuctionScraper: it communicates with the BatchManager to obtain the BatchSpecifications and it scraps the urls provided in the BatchSpecifications, then it provides the scraped data to the BatchManager so that it can create the Batch.

A **BatchSpecification** mostly provides a **Listing URL** for a **Provider** along with some characteristics for the auctions that we expect to find at that URL :

- the **family** of the auctions for that URL (mandatory)
- the **country** of the auctions for that URL (optional)
- a range of year during which the object related to the auction has been produced
  - the **start year** (optional)
  - the **end year** (optional)
- a **list of condition identifiers** for the object related to the auction (mandatory, at least one condition)
- the **URL of the last auction** that has been processed for the Listing URL (optional)

The **Provider** is a value that allows the batch extractor to better identify the web site from which it is going to extract the auctions, i.e. Delcampe, Ebay, ...

The **Listing URL** is an URL that allows to access a list of auctions who have the same characteristics as described in the batch specification. The Batch extractor will be processing each of this auctions and it will extract the following information :

- title
- image thumbnail URL
- image URL
- type (fixed price or bidding)
- starting price with the currency
- ending price with the currency (price it was sold)
- number of bids (this value is set to 1 for a fixed price auction)
- list of bids (bidder pseudo, bidder price with the currency)
- seller pseudo
- seller country

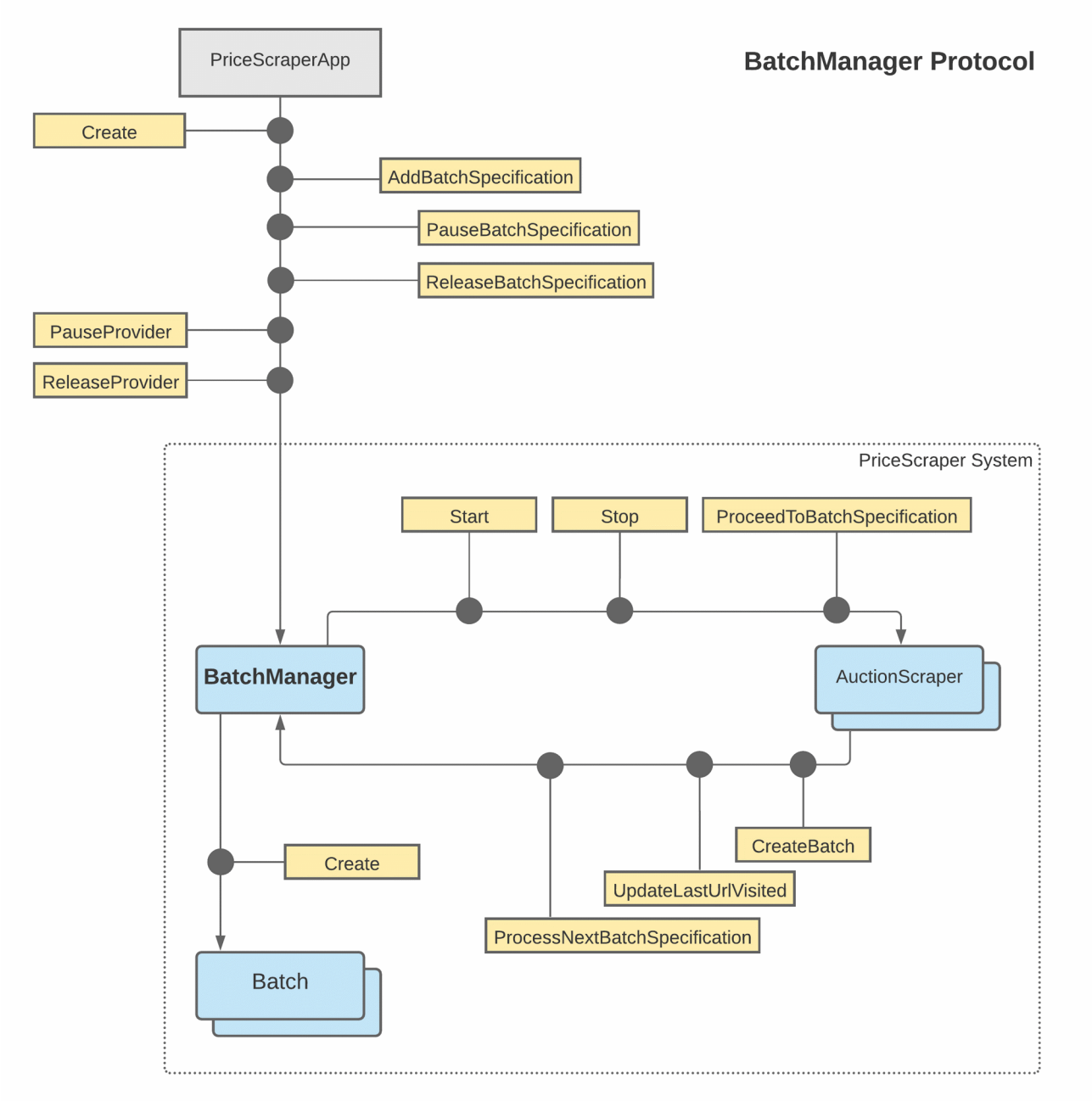
1.1. BatchManager Actor

The BatchManager is an actor that is created by the PriceScraper application when it starts and it has the following responsibilities :

- it creates and starts the AuctionScraper actors for the different providers
- it holds and manages the list of BatchSpecifications

1.1.1. Protocol

The BatchManager protocol is depicted in the schema below and detailed in the later paragraphs.



1.1.1.1. Create

After creating the BatchManager actor the PriceScraper application sends the **Create** message to the BatchManager so that it can initialize itself. During its initialization process the BatchManager creates the AuctionScraper actors for each provider and sends the **Start** message to each of this AuctionScraper actors.

1.1.1.2. AddBatchSpecification

Upon reception of this Command the BatchManager checks that no other BatchSpecification exists with the same name and adds it to the list of BatchSpecifications it holds.

1.1.1.3. PauseBatchSpecification

Upon reception of this Command the BatchManager looks for an existing BatchSpecification with the provided name and marks it as paused (a paused batch specification will not generate any new Batch, until it is released).

1.1.1.4. ReleaseBatchSpecification

Upon reception of this Command the BatchManager looks for an existing BatchSpecification with the provided name and releases it by updating the pause indicator to false, allowing this BatchSpecification to be included in upcoming scraping schedule.

1.1.1.5. PauseProvider

Upon reception of this Command the BatchManager looks for all the BatchSpecification whose provider equals the provided provider, and marks this BatchSpecifications as paused.

1.1.1.6. ReleaseProvider

Upon reception of this Command the BatchManager looks for all the BatchSpecification whose provider equals the provided provider, and releases this BatchSpecifications by updating their pause indicator to false.

1.1.1.7. Start

After creating the AuctionScraper actors the BatchManager sends the **Start** message to each AuctionScraper actor it has created (one per provider). This allows the AuctionScraper to initialize and be ready to handle other messages.

1.1.1.8. CreateBatch

Upon reception of this Command the BatchManager creates a Batch actor with the provided list of auctions and BatchSpecification details, and it sends the Create command to this Batch actor.

1.1.1.9. UpdateLastUrlVisited

Upon reception of this Command the BatchManager looks for an existing BatchSpecification with the provided BatchSpecification ID and updates the lastUrlVisited field with the provided value.

1.1.1.10. ProcessNextBatchSpecification

Upon reception of this Command the BatchManager looks for a BatchSpecification that needs to be updated and whose provider equals the provided provider. If more than one

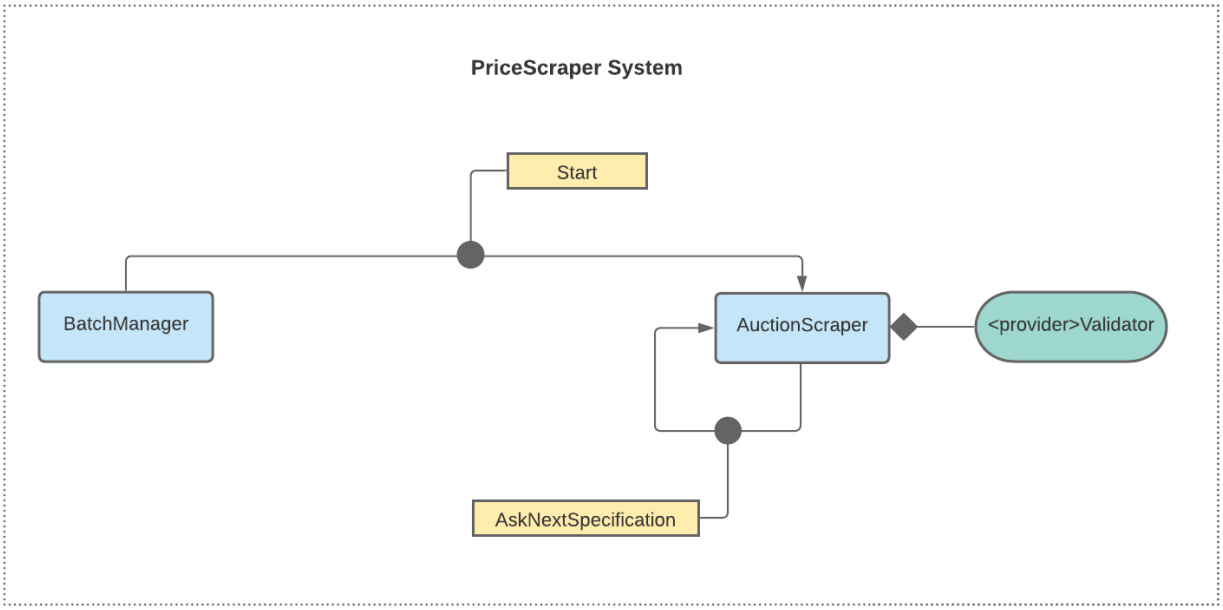
BatchSpecification meets this criteria then the **ProceedToBatchSpecification** is sent to the AuctionScraper with the BatchSpecification that has been updated the least recently.

A BatchSpecification needs to be updated if the timestamp corresponding of its updatedAt plus its intervalSeconds field is greater than the current timestamp.

1.2. AuctionScraper Actor

1.1.1. Starting the AuctionScrapers

As depicted in the schema below, each AuctionScraper has a **validator** object that is specialized in the analysis of the HTML code of a specific provider, it allows for example to fetch the auction urls from the listing page url provided in a BatchSpecification or to extract all



the information from an HTML page of an auction (auction title, auction seller, auction start price and final price, auction bids, ...).

In terms of messages, after creating the AuctionScraper actors the BatchManager sends the **Start** message to each AuctionScraper actor it has created (one per provider). This allows the AuctionScraper to initialize and move to a state where it periodically sends to itself the **AskNextSpecification** message, we will see later how this message is processed.

1.2. Process of extraction

The batch extraction process consists of the following steps :

Open the Listing URL and extract the list of auction URL
For each auction URL
...