Структура решения.

Система разбивается на некоторое число основных компонентов: Crawlers, TaskManager, Queues, Parser, Supervisor. Система может быть создана как распределенная система. Каждый модуль может быть набором микро-сервисов, за исключением супервизора. Супервизор находится на вершине иерархии, и управление всем процессом должно осуществляться через него.

Crawlers отвечает за поиск и/или индексирование страниц в сети и занесение страниц в базу данных.

TaskManager отвечает за чтение необходимых URL из базы, созданием задач на обработку, занесением данных задач в очередь. Supervisor может передавать в TaskManager новые URL, которые необходимо будет обойти. Новые URL должны будут попасть в очередь.

Queues – система очередей с задачами на обработку. Задачи из этих очередей поступают в Parser.

Parser непосредственно занимается выгрузкой страниц, сохранением новых страниц в БД. Если при разборе встретятся новые страницы, то соответствующая информация будет передана супервизору.