СОВРЕМЕННЫЕ РАСПРЕДЕЛЕННЫЕ БАЗЫ ДАННЫХ В СИСТЕМЕ АВТОМАТИЗАЦИИ ПРИНЯТИЯ РЕШЕНИЙ В УПРАВЛЕНИИ ПРЕДПРИЯТИЕМ

асп. Д.А. Скачко, Институт кибернетики им. В.М. Глушкова, г. Киев

Современные предприятия характеризуются большим количеством происходящих в них процессов, часто разнесенных географически, с большим объемом данных и жесткими требованиями к их безопасности и надежности. Все эти моменты ставят не тривиальные задачи по сбору, хранению, анализу и визуализации данных, с целью автоматизации принятия решений в процессе управления предприятием.

Современные базы данных можно разделить на реляционные СУДБ и на NoSQL решения. Каждые имеют свои преимущества и недостатки. Но учитывая динамичную составляющую современных данных как по количеству, так и по структуре, большее преимущество отдается в сторону NoSQL решений.

С точки зрения хранения данных в NoSQL базах, есть необходимость в денормализации данных, но это, в свою очередь, дает гибкость при раскладывании данных на шарды (сервера) и репликации. Но делать поиск или более сложный анализ с помощью таких решений очень не практично и медленно. Для этих целей подходят специализированные поисковые индексы, которые могут быстро индексировать данные из базы, распределять их по серверам для равномерной нагрузки и для надежности хранения самих поисковых данных. Поверх поисковой системы уже настраиваются системы анализа и визуализации, которые, в свою очередь, служат для извлечения знаний из сухих данных и для визуализации данных с целью эффективного принятия решения.

В ходе разработки был собран вычислительный кластер, который состоит из следующих частей: распределенная база данных MongoDB, с репликацией данных для надежного хранения больших объемов данных, и из поискового индекса ElasticSearch, который служит для выполнения сложных запросов по поиску и анализу данных, в том числе, и построению детальных статистик по деятельности предприятия. В завершение была установлена система Kibana для визуализации данных, поисковых и статистических запросов. Вся система работает на нескольких серверах с целью хранения объема данных, превышающего ресурсы одного сервера и распределения нагрузки как при записи данных, так и при их анализе, с целью обеспечения бесперебойной работы с дупликацией данных.