0.1. *Саклаков В.М.* Современные технологии обработки больших данных

На сегодняшний день в мире существует большое количество разнородных данных. При этом с одной стороны с каждым годом растет их количество, а с другой издержки на их поиск и интерпретацию. Данная проблемная ситуация усиливается по возрастающей от производственного к инновационному и далее к научному секторам экономики. Причем издержки в каждом из этих секторов так же возрастают при переходе от более низкого уровня обобщения к более высокому. Например, в инновационном секторе – от отдельного инновационного предприятия к инновационному кластеру и далее к инновационным системам (региональным, национальным).

Целью настоящей работы является анализ основных современных технологий обработки «больших данных» (Big Data).

В результате были выделены средства параллельной обработки огромных объемов данных разной степени структурированности: NoSQL, алгоритмы MapReduce, инструменты проекта Hadoop. Основными параметрами для Big Data являются: физический объем данных, скорость их прироста и обработки, а так же возможность одновременной обработки данных, различной степени структурированности. При этом, согласно теоремы CAP, в результате выполнения распределенных вычислениях можно обеспечить лишь два из трех свойств: согласованность данных, их доступность и устойчивость к разделению. Рассмотрим преимущества и недостатки вышеописанных средств обработки:

1. NoSQL. Применяется для реализации хранилищ БД, до определенной степени решая проблему масштабируемости. С одной стороны данные средства обработки гарантируют завершение каждого запроса (успешное или нет), производят изменения системы с целью согласования данных (уже существующих или новых), обеспечивают конечную согласованность данных. С другой стороны именно последней приходится жертвовать проектировщикам NoSQL-систем для обеспечения доступности данных и их устойчивости к разделению. Примерами СУБД данного класса могут служить Apache Cassandra, MongoDB.

2. MapReduce. Область применения данной модели – вычисление распределенных задач в компьютерном кластере. Основным преимуществом данного подхода является производить обработку и свертку данных параллельно и независимо друг от друга на различных серверах. При этом сортировка петабайта данных может занять лишь несколько часов. Недостатком является тот факт, что данный алгоритм в своей эффективности может уступать более последовательным алгоритмам.

3. Hadoop. Область применения данного комплекса – реализация поисковых и контекстных механизмов в веб-сервисах с высокой нагрузкой. Основным преимуществом Hadoop является масштабируемость кластера по горизонтальному типу с применением общедоступного оборудования, вместо мощных и дорогостоящих серверов. Однако масштабируемость все же имеет ограничение по количеству узлов. Другим узким местом данного подхода является размер оперативной памяти на каждом узле.

В рамках последующей научной работы планируется проводить более детальный анализ технологий обработки Big Data, а так же существующих подходов к разработке гетерогенных хранилищ данных.

Научный руководитель – к.т.н. Иванов М. А.