ИПИ РАН

Системы высокой доступности и Большие Данные

Заместитель директора по научной работе
Института проблем информатики РАН
Доктор технических наук
Член-корреспондент Академии криптографии РФ

Будзко Владимир Игоревич

Высокая доступность означает, что любой пользователь может обратиться к автоматизированной системе и получить доступ в рамках своих полномочий к необходимым ему ресурсам и услугам за приемлемое для него (пользователя) время. Причем, приемлемое время доступа сохраняется не только при существующих нагрузках, но и при ожидаемом их увеличении на заданную перспективу, деградации архитектуры системы, вызванной отказами, массовым уничтожением ее компонентов и (или) целых объектовых комплексов и связей между ними.

Высокая доступность

- должна быть достаточной при ожидаемом расширении нагрузок на заданную перспективу
- доступность должна сохраняться и при особых условиях
- минимальная высокая доступность = 99,9%

Высокая доступность автоматизированной информационной системы (АИС) требуется, прежде всего, когда использование результатов работы АИС - неотъемлемая часть технологии функционирования самой ОРГАНИЗАЦИИ и прерывание работы АИС нарушает непрерывность бизнес-процесса. Критические направления деятельности особенно требовательны к доступности.

Высокая доступность:

Доступность определяется по формуле

$$\text{СВНО}/(\text{СВНО} + \text{СВВ})100\% \leq 99{,}9\% \quad ,$$

где:

СВНО (среднее время наработки на отказ) – среднее время, которое система работает без отказов после установки и запуска или восстановления;

СВВ (среднее время восстановления) – среднее время восстановления после отказа.

СВД+Big Data=СВДД

- СВД, которая работает в среде Big Data, должна обладать вторым важным свойством – доступность всех необходимых данных для своевременной выработки адекватного информационного продукта, на основе которого может быть принято оптимальное решение.
- От полноты и точности доступных исходных данных, зависит качество получаемого «информационного продукта» (ИП) и соответственно качество информационной поддержки процесса принятия решений пользователем.
- средства своевременного сбора точных и полных данных и средства их своевременной обработки для получения информации, обеспечивающей своевременное принятие эффективного решения

СВД+Big Data=СВДД

- АИС ВД должна включать средства своевременного сбора точных и полных данных и средства их своевременной обработки для получения информации, обеспечивающей своевременное принятие эффективного решения.
- Поэтому уместно называть такие системы не просто СВД, а системами высокой доступности данных – СВДД

Новый взгляд – старые принципы

- В какой степени доступные данные отражают реальное состояние моделируемой предметной области? ***Полнота.***
- Насколько правильно данные описывают предметную область? ***Точность.***
- ***Система высокой доступности данных***: доступные данные достаточной полноты и точности обработаны и вовремя получен информационный продукт (ИП).

выявление (***D***iscovery),

↓

отбор (***D***iscrimination),

↓

переработка (***D***istillation),

↓

доведение в нужном представлении (***D***elivery/***D***issemination).

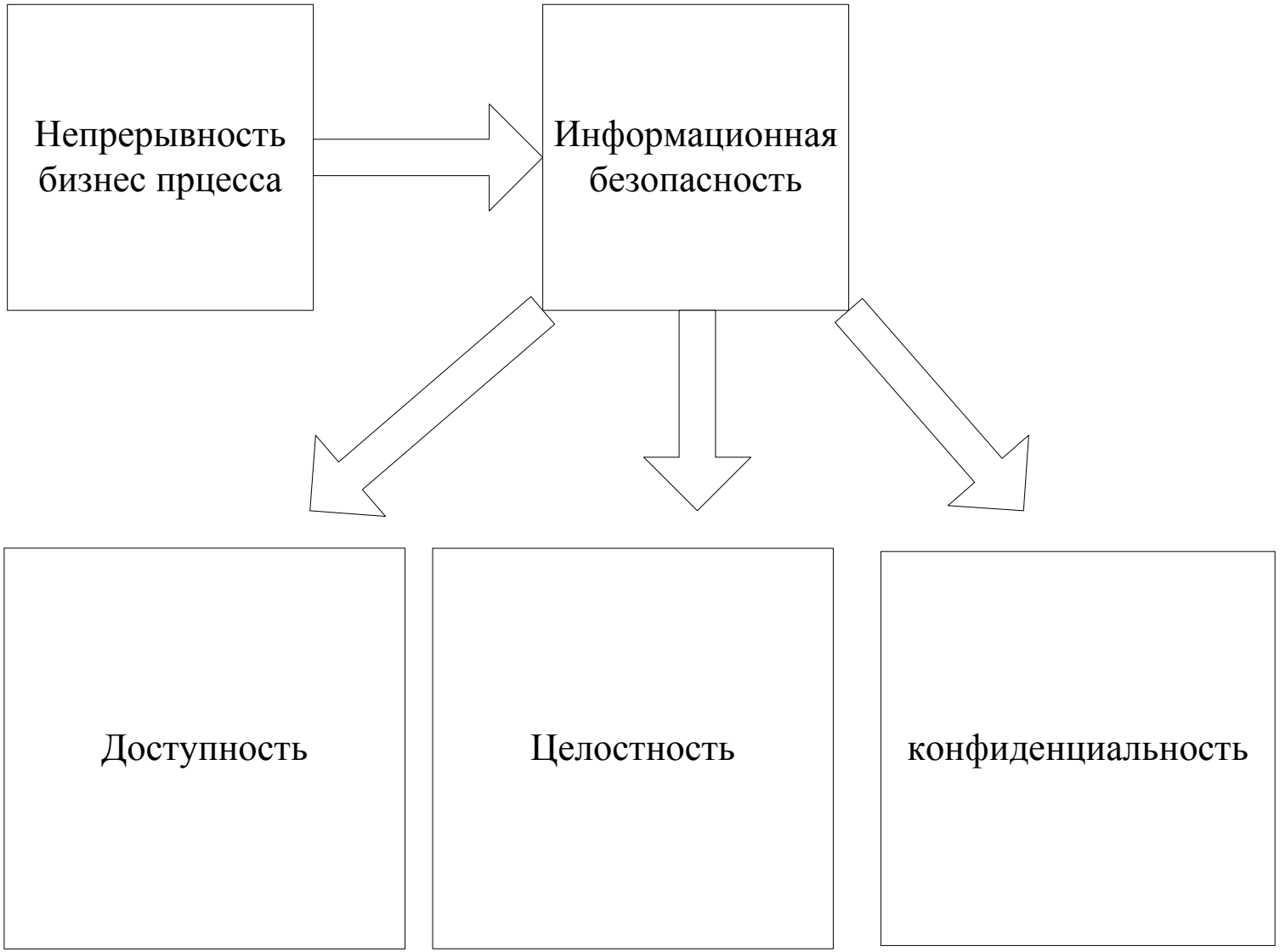
→

Т - время, прошедшее с момента возникновения релевантных исходных данных (ИД) до момента завершения выполнение действия на основании выработанного ИП

$$T = T_В + T_П + T_О + T_{ИП} + T_Р + T_Д$$

$$T < T_{РЕГ}$$

Основа обеспечения непрерывности бизнес процесса – безопасность информации



Критические точки смысловой обработки СВДД Big Data

Кто	Что	доступность	целостность	конфиденциальность	полнота	точность
Поисковик	**ИД обработчику**	**Да**	**Да**	**?**	**Да**	**Да**
Обработчик	**данные аналитику**	**Да**	**Да**	**?**	**Да**	**Да**
Аналитик	**подготовка ИП**	**Да**	**Да**	**?**	**–**	**–**
Руководитель	**решение**	**Да**	**Да**	**?**	**–**	**–**
Объект управления	**действие**	**Да**	**–**	**–**	**–**	**–**

Компетенция ИПИ РАН

Технологические решения

- извлечение данных из больших коллекций неструктурированных и разно-структурированных данных;
- структуризация извлеченных данных, преобразование и чистка, обеспечение их достоверности, семантическая интеграция при получении из различных коллекций; формирование хранилища данных для дальнейшего решения конкретных аналитических задач;
- анализа данных в среде map-reduce с использованием свободно распространяемого ПО с открытым кодом (например, язык R), так и фирменного ПО (например, комплекс SPSS от IBM), предоставляющих методы статистического анализа и машинного обучения

Технологические решения

- поддержка аналитики со сложными методами анализа данных (машинное обучение и статистика в IBM Warehouse или IBM Netezza) на машинах баз данных, без передачи данных по сети;
- обработка потоковых данных на лету, в реальном масштабе времени со временем отклика порядка миллисекунд (например, система обработки потоковых данных IBM Streams);
- организация крупных хранилищ структурированных данных на основе параллельных машин баз данных при использовании ETL техники;
- поддержка виртуальной интеграции данных для создания федеративных баз данных (например, в IBM Federation Server);

Технологические решения

- создание облачных инфраструктур со встроенной Hadoop и развитой ETL технологией для организации крупных хранилищ данных и средств анализа данных над ними;
- поддержка развитых методов и средств анализа данных, включающих:
  - ✓ статистический анализ данных,
  - ✓ интеллектуальный анализ данных,
  - ✓ анализ текстов, в том числе синтаксический разбор и смысловое аннотирование текста, извлечение структурированной информации из текста.

Cloud computing для Big Data
Точки зрения на «круглых столах»

- Фактически, cloud computing – это возвращение эпохи мейнфреймов – гигантских суперкомпьютеров
- Достоинства cloud computing – снижаются требования к вычислительной мощности ПК (непременным условием является только наличие доступа в Интернет);
  – отказоустойчивость;
  – безопасность;
  – высокая скорость обработки данных;
  – снижение затрат на аппаратное и программное обеспечение, на обслуживание и электроэнергию;
  – экономия дискового пространства (и данные, и программы хранятся в Интернете).
- Недостатки cloud computing:
  – зависимость сохранности пользовательских данных от компаний, предоставляющих услугу cloud computing;
  – появление новых («облачных») монополистов

Спасибо за внимание

Заместитель директора по научной работе
Института проблем информатики РАН
Доктор технических наук
Член-корреспондент Академии
криптографии РФ

Будзко Владимир Игоревич

vbudzko@ipiran.ru

ИПИ РАН