



Методические аспекты обучения специалистов по обработке и анализу данных с использованием Data GRID в среде виртуальной компьютерной лаборатории

Belov M.A., Cheremisina E.N., Tokareva N.A.

NEC'2019, Montenegro

ИНСТИТУТ СИСТЕМОГО АНАЛИЗА И УПРАВЛЕНИЯ



- ✓ Важные аспекты в подготовке специалистов по обработке и анализу данных (Data Scientists)
- ✓ Виртуальная компьютерная лаборатория (ВКЛ) в Институте Системного Анализа и Управления (ИСАУ) государственного университета «Дубна»
- ✓ Организация обучения в среде ВКЛ навыкам проектирования и развертывания кластеров Data Grid, администрирования и выполнения задач по обработке реальных данных
- ✓ Выводы

Важные аспекты в подготовке специалистов Data Scientists

Профессия Data Scientist — специалист по обработке, анализу и хранению больших массивов данных Big Data— в современном мире считается одной из самых перспективных, актуальных и высокооплачиваемых.

Определяющие факторы подготовки Data Scientist:

- хорошая математическая подготовка;
- **освоение конкретных технологий** (например: SAS, Hadoop, язык R или решения IBM);
- активно вовлечение в решение прикладных задач для Big Data.

Data GRID – концепция

- Data GRID - это общий термин для обозначения нескольких сайтов или кластеров для распределенной обработки и хранения данных.
- HDFS Hadoop - это метод или способ реализации Data GRID для хранения данных, многие другие инфраструктуры Hadoop, такие как MapReduce или Spark, используются для распределенной обработки данных.
- В GRID - компьютеринге компьютеры в сети могут совместно работать над задачей, функционируя как суперкомпьютер. Другой взгляд: GRID-вычисления теперь являются традиционной высокопроизводительной системой с разновидностью MPI.
- Мы рассматриваем GRID как концепцию распределенной системы - способ использования компьютеров, распределенных по сети, для решения проблемы. GRID - это группа физических машин, соединенных для создания компьютера GRID, а Hadoop - это программное обеспечение, работающее на этих машинах, поэтому Hadoop является подмножеством Grid-вычислений.

Сложности в обеспечении процесса освоения технологий (в том числе для студентов ЦДЗО ИСАУ)

- Быстрое изменение технологий.
- Обеспечение для студентов практического опыта развертывания распределенных многокомпонентных систем.
- Недостаточное количество аудиторных часов того, чтобы выполнить полный набор практических упражнений, которые помогают студентам изучать сложные информационные системы.
- На типичном ПК со средними возможностями невозможно получить реальный практический опыт работы с многокомпонентным кластером Data GRID (требования к оборудованию для таких систем выходят за рамки того, что предлагается на типичных ПК).
- Высокая стоимость однопользовательской лицензии на некоторые программные компоненты или профессиональную техподдержку.

ВКЛ ИСАУ как средство решения указанных проблем

Виртуальная Компьютерная Лаборатория (ВКЛ) - комплекс программно-аппаратных средств виртуализации и контейнеризации.

Возможности:

- гибко и по требованию предоставлять и использовать вычислительные ресурсы в виде облачных интернет-сервисов;
- имеет интегрированную систему управления знаниями, основанную на принципах самоорганизации;
- функционируют как однородная среда с элементами когнитивного представления внутренних операционных ресурсов на основе визуальных моделей;
- позволяет выполнять ресурсоемкие вычислительные расчеты и задачи, связанные с разработкой сложных корпоративных и других распределенных информационных систем.
- предоставляет выделенные виртуальные серверы для инновационных проектов.

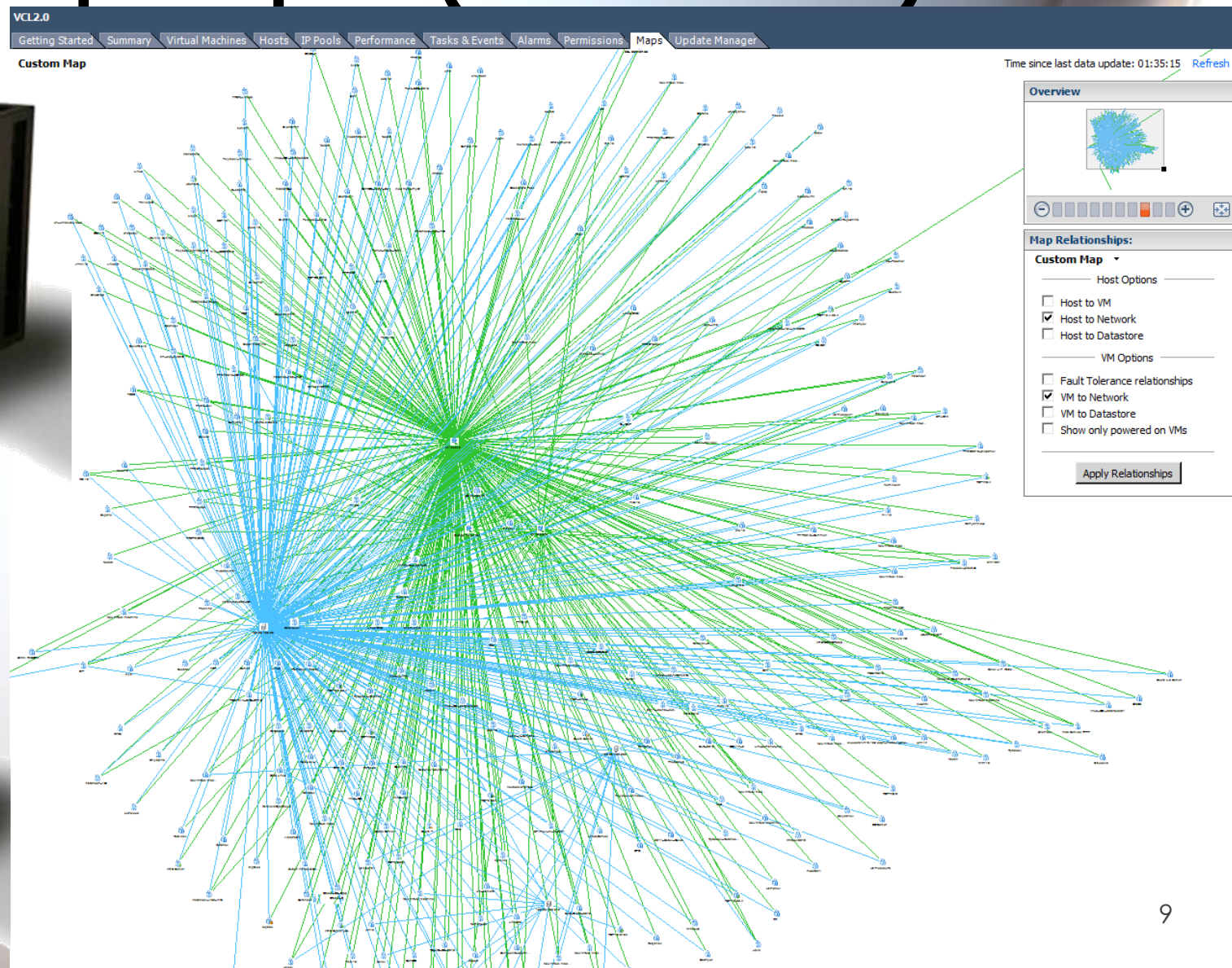
Технические вызовы для ВКЛ при обучении Data GRID

- Увеличение требований к хранению
- Несколько поставщиков
- Сложные системы
- Устойчивый рост количества пользователей
- Консолидация серверов
- Временные ограничения
- Надежность оборудования
- Сложное администрирование

Ключевые компоненты ВКЛ

- **VMware vSphere & Docker** – технологии виртуализации и контейнеризации;
- **METALOLOM** – Multiuser Enhanced Training And Learning Operations Lights Out Manager (системное администрирование, сервис менеджмент, управление обучением)
- **CPESDA** – Cloud Personal Expert System Deployment Assistant (основанная на облачных технологиях система, позволяющая студентам автоматизировать развертывание сложного программного обеспечения)
- **AIKMS** – Artificial Intelligence Knowledge Management System (KMS, использующая машинное обучение для анализа текста, распознавания и обработки речи, а также для интеллектуальных чат-ботов)

Сетевой сегмент Виртуальной компьютерной лаборатории (май 2019 г.)



Примеры задач, решаемых с использованием ВКЛ

1

Проектирование и развертывание кластера Data GRID на основе программного обеспечения Apache Hadoop с использованием наиболее распространенных топологий:

- базовая горизонтальная топология,
- топология федерации,
- монадическая топология,
- иерархическая топология,
- гибридная топология

Примеры задач, решаемых с использованием ВКЛ

2

Выполнение основных задач администрирования кластера:

- добавление или удаление хостов и экземпляров службы,
- изменение коэффициента репликации,
- настройка объема выделенной памяти для контейнеров выполнения
- и т. д.

Примеры задач, решаемых с использованием ВКЛ

3

Case Study в области DataGRID

- Загрузить реальные данные из различных источников данных в распределенную файловую систему HDFS и выполнить перебалансировку данных.
- На основе загруженных данных и аналитических и служебных инструментов (MapReduce, Spark, HUE, HCatalog, Hive, Impala, Pig Latin, Sqoop, Solr, Oozie, CDSW) выполнить анализ, например, основываясь на нескольких десятках миллионов постов с технического форума, оценить:
 - популярность языков программирования,
 - эффективность модерации,
 - тональность высказывания по данному товару ,
 - и т. д.

Управление знаниями в среде ВКЛ

Технологические инструменты, которые могут использоваться для передачи формальных знаний, включают:

- базы знаний с различными функциями поиска;
- блоги, вики и социальные сети;
- «Вики-учебники», которые позволяют всем участникам совместно создавать и обновлять образовательный контент и обмениваться практическими задачами (**в том числе от реальных компаний**);
- блоги пользователей, форумы и системы групповых чатов, интеллектуальные чат-боты.

Выводы

Методические аспекты подготовки специалистов Data Scientist включают:

- Обучение созданию и использованию Data Grid в среде Виртуальной Компьютерной Лаборатории ИСАУ с использованием наиболее популярных инструментов из инфраструктуры Apache Hadoop.
- Внедрение самостоятельных практических занятий с использованием ВКЛ как для студентов очной, так и заочной формы обучения с применением дистанционных технологий.
- Совместную работу в системе управления знаниями, к которой могут быть подключены потенциальные работодатели.

Это позволяет повысить качество подготовки специалистов по обработке и анализу данных в ИСАУ университета «Дубна».

СПАСИБО ЗА ВНИМАНИЕ!