**4. Експериментальний розділ**

В попередніх розділах даної МКР виконувалося дослідження можливості побудови конфігурованого сховища даних на базі розподіленої платформи Hadoop та проектування програмного рішення, на основі виконаних досліджень. Система була спроектована і побудована з урахуванням всіх основних вимог що ставляться до типових сховищ даних, а також з підтримкою легкої конфігурації системи в залежності від предметної області, в якій вона буде використовуватися. В даному розділі виконується перевірка коректності роботи системи та її функціональне тестування.

Тестування продуктивності системи не входить в межі даної роботи і потребує наявності достатньо великих обсягів експерементальних даних та відповідної інфраструктури (кластера великої кількості машин). Таке тестування повинне проводитися за наявності деталізованих вимог до сховища даних та з урахуванням інфраструктурних можливостей. Майже лінійна масштабованість платформи Hadoop вже була підтверджена в лабораторних та реальних умовах і є описана в багатьох наукових доповідях.

Даний розділ покриває тільки функціональне тестування системи і виконує валідацію лише базового функціоналу, що є характерним для усіх сховищ даних, незалежно від предметної області. Такий функціонал включає наступні елементи:

* Коректне завантаження структурованих даних з СКБД в розподілену файлову систему;
* Коректне завантаження окремих файлів з файлових серверів в розподілену файлову систему;
* Побудова таблиць фактів з заданої множини вхідних даних;
* Побудова таблиць вимірів різних типів з заданої множини вхідних даних;
* Надання інтерфейсу до підготованого сховища даних у вигляді, зручному для використання споживачами даних.

Оскільки розроблене рішення є фреймворком і не містить в собі ніяких вбудованих правил обробки, окрім як правил формування таблиць вимірів різних типів, то характер даних для виконання тестування не є важливим. Основною вимогою до цих даних є наявність таких полів:

* Дата появи запису;
* Унікальний ідентифікатор запису;
* Значення яке несе цінність в процесі опрацювання даних(значення яке використовується для побудови аналітичного звіту).

Для валідації системи було згенеровано штучні дані, які відповідають всім вищевказаним вимогам. Кількість тестових даних не є великою, адже вони необхідні тільки для перевірки функціональної правильності системи.

Набір тестових даних складається з чотирьох таблиць наповнених штучно згенерованими даними. В кожній таблиці містяться дані, що використовуватимуться для побудови окремої логічної частини вітрини даних: таблиця фактів, таблиці вимірів першого, другого та третього типів. Ідентичні дані містяться в текстових файлах на локальній файловій системі, яка виконує роль файлового сервера. Для легкості тестування дані були вибрані ідентичними. Записи збережені у форматі розділеному вертикальними лініями (‘|’). Кожне поле в записі відділяється від сусіднього вертикальною лінією. Поширеним форматом є також розділяння комами, але такий спосіб не завжди підходить, адже інколи значення полів можуть містити символи коми. Система сховища даних буде розгорнута у віртуальному середовищі VirtualBox, з використанням уже готового, доступного у вільному доступі, віртуального образа Hortonworks HDP 2.2 Sandbox. База даних з тестовими таблицями працює на хості, засобами якого виконується віртуальне середовище. Для тестових цілей була обрана open-source СКБД MySQL. Текстові файли з тестовими даних знаходяться в локальній файловій системі того ж хоста, на якому працює СКБД.

Запуски виконуються шляхом надсилання послідовностей виконання (описаних в форматі xml, з урахуванням вимог, що ставляться засобом координації виконання Oozie) через командну стрічку на сервер-координатор. Послідовності виконання повинні знаходиться в розподіленій файловій системі(HDFS), а конфігураційний файл з параметрами виконання – на локальній файловій системі, звідки виконується запуск.