这是一份关于 **“赋能你的 AI 智能体：在 Cassandra 中探索向量搜索 (Empowering Your AI Agents: Exploring Vector Search in Cassandra)”** 讲座/研讨会的音频总结。

这次讲座由 IBM Watsonx Data / Cassandra 团队的 Aaron 和 Patrick 主持，主要通过一个名为 **“Killer Video”**（一个类似 YouTube 的视频应用）的实战案例，演示了如何利用 **Apache Cassandra (Astra DB)** 的向量搜索功能来构建生成式 AI 应用。

以下是本次讲座的核心内容总结：

1. 背景与概念介绍

- **背景：** 演讲者提到 IBM 收购 DataStax 后，整合了 Cassandra 的能力。Cassandra 被广泛用于全球大型企业的关键业务中。
- **生成式 AI 架构 (RAG)：**
    - 传统的 AI 应用流程是：用户输入 -> LLM（大语言模型）-> 返回结果。
    - **演讲者强调的架构 (RAG - 检索增强生成)：** 在用户与 LLM 交互之前，先通过数据库（Cassandra）进行**向量搜索**，检索相关的上下文数据（Context），再将其与用户提示词一起发送给 LLM。这需要一个离线的向量嵌入（Embedding）过程。
- **什么是向量 (Vector)？** Patrick 解释了向量是具有方向和大小的数值数组。**向量嵌入 (Embeddings)** 将文本、视频或音频转化为浮点数数组。
- **搜索机制：** 传统的数据库使用关键词匹配，而向量数据库使用 **ANN（近似最近邻）**算法。它通过计算向量之间的距离（如余弦相似度）来寻找“语义上相似”的内容（例如：搜索“80年代空手道电影”能找到《龙威小子》，即使没有完全匹配的关键词）。

2. 研讨会实战演练 (Hands-on Lab)

讲座的核心部分是指导听众完成一个具体的开发练习。目标是为“Killer Video”应用构建一个**“相关视频推荐”**功能。

步骤一：创建数据库

- 使用 **Astra DB**（DataStax 的托管 Cassandra 服务）。
- 创建一个 Serverless 向量数据库，选择云服务商（如 Google Cloud）和地区。

步骤二：定义 Schema (表结构)

- 在数据库中创建一个名为 `videos` 的表。

- 关键点： 表中包含一个特定的列用于存储向量数据。在本次练习中，向量维度被设定为 **384**（对应所使用的 IBM Granite embedding model 或类似的 HuggingFace 模型）。
- 创建索引： 使用 **SAI (Storage Attached Indexing)** 创建自定义索引，并指定使用余弦相似度 **(Cosine Similarity)** 作为度量标准。这是进行高效向量搜索的关键。

步骤三：加载数据

- 使用 Python 脚本加载数据。
- 为了节省研讨会时间，他们没有实时生成嵌入（Embedding），而是使用了一个预先处理好的 CSV 文件，其中包含 500 个视频及其对应的向量数据。
- 使用了 `cassandra-driver` 来连接数据库并插入数据。

步骤四：后端实现与查询

- 环境： 使用 Python 和 FastAPI 构建后端。
- 连接配置： 需要下载 `Secure Connect Bundle` (SCB) 和创建 Application Token 来进行安全连接。
- 执行向量搜索： 演示了如何编写 SQL/CQL 查询，利用向量索引查找与当前视频最相似的其他视频。

步骤五：前端集成

- 演示了一个 React 前端应用。
- 当用户点击一个视频时，前端调用后端 API，后端在 Cassandra 中进行向量搜索，返回“相关视频”列表并展示在侧边栏。

3. 关键技术栈

- 数据库： Astra DB (基于 Apache Cassandra)。
- 语言： Python (后端脚本), JavaScript/React (前端)。
- 库/工具： `cassandra-driver`, `fastapi`, `numpy` (用于处理数组), HuggingFace (用于模型/dataset)。

4. 总结与后续

- 讲座展示了如何从零开始构建一个具备语义搜索功能的应用，而不仅仅是简单的关键词匹配。

- 演讲者最后提到了 GitHub 上的代码库供大家后续参考，并宣传了即将重启的 Cassandra 用户组（User Groups）活动（如在纽约、明尼阿波利斯等地）。

简而言之，这是一个技术导向的讲座，旨在教开发者如何利用 **Cassandra 的向量搜索能力**来为应用程序添加类似“猜你喜欢”或“相关推荐”的 AI 功能。