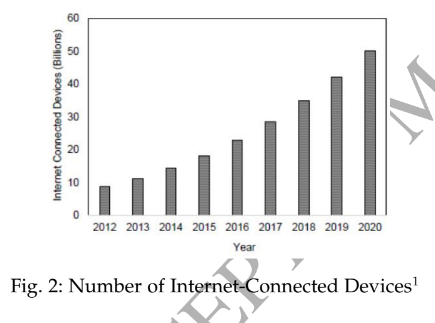
**Vai trò của phân tích dữ liệu lớn trong Internet vạn vật**

**Tóm tắt** - sự tăng trưởng bùng nổ về số lượng thiết bị kết nối với Internet vạn vật (IoT) và việc tiêu thụ dữ liệu tăng theo cấp số nhân đã cho thấy sự tăng trưởng của dữ liệu lớn hoàn toàn trùng lặp với mức tăng trưởng của IoT. Việc quản lý dữ liệu lớn trong một hệ thống mạng liên tục mở rộng làm phát sinh những lo ngại không hề nhỏ về hiệu quả thu thập dữ liệu, dữ liệu xử lý, phân tích và bảo mật. Để giải quyết những lo ngại này, các nhà nghiên cứu đã kiểm tra các thách thức liên quan với việc triển khai thành công IoT. Mặc dù có rất nhiều các nghiên cứu về dữ liệu lớn, IoT, phân tích, sự hội tụ của những lĩnh vực trên đã tạo ra cơ hội để phát triển dữ liệu lớn, phân tích dữ liệu trong IoT. Trong bài báo này, chúng tôi sẽ khám phá những tiến bộ gần đây trong phân tích dữ liệu lớn cho các hệ thống IoT cũng như các yêu cầu chính để quản lý dữ liệu lớn cũng như tạo điều kiện phân tích dữ liệu trong môi trường IoT. Chúng tôi phân loại tài liệu dựa trên những thông số quan trọng, xác định các cơ hội do sự hội tụ của dữ liệu lớn, phân tích và IoT cũng như thảo luận về vai trò của phân tích dữ liệu lớn trong các ứng dụng IoT. Cuối cùng, một số thách thức mở được trình bày như là hướng nghiên cứu trong tương lai.

**Chỉ** **mục** - Internet vạn vật, dữ liệu lớn, điện toán phân tán, thành phố thông minh

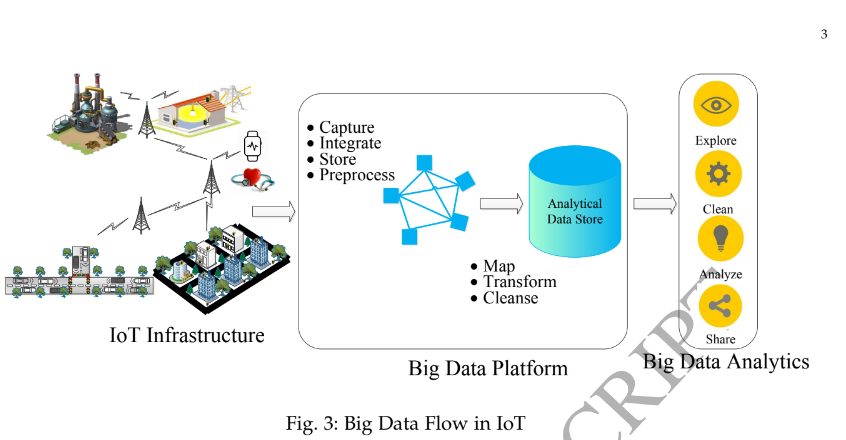
1. **Giới thiệu**

Những tiến bộ công nghệ và sự hội tụ nhanh chóng của truyền thông không dây, kĩ thuật số điện tử và hệ thống cơ điện vi mô đã dẫn đến sự xuất hiện của IoT. Theo báo cáo của Cisco, số lượng thiết bị kết nối internet đã vượt quá dân số toàn cầu. Những đối tượng kết nối internet này, bao gồm PC, điện thoại thông minh, cảm biến hỗ trợ wifi, thiết bị đeo được và đồ gia dụng, hình thành nên IoT như trong hình 1. Báo cáo cho thấy số lượng thiết bị kết nối Internet sẽ tăng gấp đôi từ 22.9 tỷ vào năm 2016 lên 50 tỷ vào năm 2020 như trong hình 2.



Hầu hết các ứng dụng IoT không chỉ tập trung vào việc giám sát các sự kiện riêng biệt mà còn khia thác thông tin được thu thập bởi các thiết bị IoT. Phần lớn thiết bị thu thập dữ liệu trong môi trường IoT đều được trang bị cảm biến yêu cầu các giao thức tùy chỉnh, chẳng hạn như vận chuyển hàng đợi tin nhắn từ xa, dịch vụ phân phối dữ liệu. Giả dụ rằng các cảm biến được sử dụng trong hầy hết các ngành công nghiệp, IoT dự kiến sẽ tạo ra một lượng lớn dữ liệu. Dữ liệu được tạo ra từ các thiết bị IoT có thể được dùng trong việc tìm kiếm các xu hướng nghiên cứu tiềm năng và điều tra sự tác động của các sự kiện hoặc các quyết định. Những dữ liệu này được xử lý bằng các công cụ phân tích khác nhau.

Hình 3 minh họa quá trình thu thập, giám sát và phân tích dữ liệu. Mặc dù IoT đã tạo ra những cơ hội chưa từng có để tăng doanh thu, giảm chi phí và cải thiện hiệu năng, nhưng việc thu thập một lượng lớn dữ liệu là không cần thiết. Để tạo doanh thu từ IoT, doanh nghiệp cần tạo ra một nền tảng nơi họ có thể thu thập, quản lý và phân tích một khổi lượng lớn dữ liệu theo một phương án có khả năng mở rộng và hiệu quả về chi phí. Trong bối cảnh này, tận dụng một nền tảng dữ liệu lớn và có thể hỗ trợ tiêu thụ, đọc nguồn dữ liệu đa dạng cũng như tăng tốc quá trình tích hợp dữ liệu trở nên rất quan trọng. Dữ liệu tích hợp và phân tích cho phép các tổ chức cách mạng hóa quy trình kinh doanh của họ. Cụ thể, các doanh nghiệp này có thể sử dụng các công cụ phân tích dữ liệu để chuyển đổi một khối lượng lớn dữ liệu được thu thập cảm biến thành những hiểu biết có giá trị. Với các xu hướng nghiên cứu chồng chéo trong các lĩnh vực này, bài viết này tập trung vào những tiến bộ gần đây trong việc quản lý dữ liệu lớn và phân tích trong mô hình IoT.



Những đóng góp của báo cáo này như sau:

• Chúng tôi phê bình nghiêm túc các tài liệu gần đây.

• Chúng tôi thảo luận về xử lý dữ liệu lớn và các nền tảng phân tích trong môi trường IoT.

• Chúng tôi thảo luận về các yêu cầu chính để xử lý và phân tích dữ liệu lớn trong môi trường IoT.

• Chúng tôi phân loại tài liệu dựa trên các thông số quan trọng.

• Chúng tôi thảo luận về các cơ hội tiềm năng trong xử lý và phân tích dữ liệu lớn trong mô hình IoT và nêu bật vai trò của phân tích dữ liệu trong các ứng dụng IoT.

• Chúng tôi thảo luận về những thách thức nghiên cứu mở và làm nổi bật tầm nhìn của các phân tích dữ liệu lớn trong IoT như là hướng nghiên cứu trong tương lai.

Các đóng góp trên sẽ được đưa ra ở phần 2-8. Phần 9 sẽ là phần tổng kết.

1. **Những cải tiến gần đây trong lĩnh vực dữ liệu lớn và phân tích dữ liệu dựa trên IoT**

Bashir và Gill đề xuất một khung phân tích dữ liệu lớn IoT để vượt qua các thách thức về lưu trữ và phân tích lượng lớn dữ liệu có nguồn gốc từ các tòa nhà thông minh. Khung đề xuất bao gồm ba thành phần là quản lý dữ liệu lớn, cảm biến IoT và phân tích dữ liệu. Các phân tích được thực hiện trong thời gian thực để được sử dụng trong các phần khác nhau của tòa nhà thông minh để quản lý mức oxy, khói / khí độc hại và độ sáng. Khung được triển khai trong phân phối Cloudera Hadoop – nơi việc phân tích được thể hiện bằng Pyspark. Kết quả cho thấy khung có thể được sử dụng để phân tích dữ liệu lớn có hỗ trợ IoT. Khung đề xuất được thiết kế đặc biệt cho các tòa nhà thông minh nên được mở rộng để làm cho nó trở nên khái quát để có thể hoạt động với các ứng dụng IoT khác bao gồm thành phố thông minh và máy bay thông minh.

C.Lee và đồng nghiệp đề xuất một hệ thống vật lý không gian mạng dựa trên IoT hỗ trợ phân tích thông tin và phương pháp thu nhận kiến thức để cải thiện năng suất trong các ngành công nghiệp khác nhau. Hệ thống này, tập trung vào phân tích dữ liệu lớn công nghiệp, tích hợp các thành phần phân tích dữ liệu khác nhau dưới dạng các mô đun có thể cấu hình lại và có thể thay thế để đáp ứng các nhu cầu kinh doanh khác nhau. Các tác giả cũng cung cấp một khung thông tin bối cảnh mới có thể giúp xử lý tin học công nghiệp dựa trên các yếu tố, địa điểm và dữ liệu phi cấu trúc để khai thác dữ liệu lớn. Một nghiên cứu trường hợp cũng được thực hiện để minh họa thiết kế của hệ thống vật lý không gian mạng được đề xuất.

P. Rizwan và cộng sự nghiên cứu những điểm mạnh và điểm yếu của các hệ thống quản lý giao thông khác nhau. Họ đề xuất một hệ thống quản lý lưu lượng thời gian thực, chi phí thấp, triển khai các thiết bị và cảm biến IoT để nắm bắt thông tin giao thông thời gian thực. Cụ thể, các cảm biến phát hiện giao thông chi phí thấp được nhúng ở giữa đường cho mỗi 500 hoặc 1000 mét. Dữ liệu được thu thập sau đó được gửi đến các công cụ phân tích để phân tích mật độ lưu lượng và cung cấp giải pháp thông qua phân tích dự đoán. So với các hệ thống hiện có, hệ thống được đề xuất cung cấp một phương pháp thay thế tốt hơn để quản lý lưu lượng.

Q. Zhang và cộng sự đề xuất Firework, một mô hình điện toán mới cho phép xử lý và chia sẻ dữ liệu phân tán trong môi trường cạnh hợp tác dựa trên IoT. Firework kết hợp dữ liệu được phân phối vật lý bằng cách cung cấp chế độ xem dữ liệu ảo cho người dùng cuối bằng các giao diện được xác định trước. Các giao diện này có dạng một tập hợp các hàm và một bộ dữ liệu. Firework hướng tới giảm thiểu độ trễ truy cập dữ liệu bằng cách di chuyển quá trình xử lý đến gần các trình cung cấp dữ liệu trong mạng biên. Ví dụ pháo hoa có nhiều bên liên quan, những người phải đăng ký bộ dữ liệu của họ và các chức năng tương ứng được trừu tượng hóa dưới dạng xem dữ liệu. Các chế độ xem dữ liệu này có sẵn cho tất cả những người tham gia của cùng một thể hiện khung để họ có thể hợp nhất nhiều chế độ xem dữ liệu vào một công việc để thực hiện phân tích dữ liệu chi tiết. Họ minh họa khái niệm đó bằng cách thực hiện các nghiên về sức khỏe được kết nối và tìm thấy những gì đã mất.

M. M. Rathore và cộng sự đề xuất một hệ thống quản lý thành phố thông minh dựa trên IoT khai thác và phân tích dữ liệu lớn. Dữ liệu được thu thập bằng cách triển khai các cảm biến khác nhau, bao gồm cảm biến thời tiết và nước, cảm biến mạng lưới, đối tượng giám sát, cảm biến căn hộ thông minh và cảm biến đỗ xe thông minh. Một kiến trúc cho hệ thống và mô hình để thực hiện cũng đã được thiết kế. Hệ thống đề xuất được triển khai bằng hệ sinh thái MapReduce Hadoop trong một môi trường thực sự. Quá trình thực hiện bao gồm một số bước, bao gồm tạo dữ liệu, thu thập dữ liệu, kết hợp dữ liệu, phân loại dữ liệu, tiền xử lý dữ liệu và ra quyết định. Spark thay thế Hadoop được sử dụng để xử lý hiệu quả dữ liệu lớn. Các hệ thống thông minh được sử dụng làm nguồn dữ liệu của thành phố để phát triển thành phố thông minh như một hệ thống được triển khai. Tuy nhiên, hệ thống thông minh được phát triển vẫn chưa được triển khai và độ chính xác của nó vẫn chưa được kiểm chứng.

B. Ahlgren và cộng sự thảo luận về tầm quan trọng của việc sử dụng IoT để cung cấp các dịch vụ nhằm cải thiện cuộc sống của người dân, bao gồm vận chuyển, chất lượng không khí và hiệu quả năng lượng. Các chuyên gia nhấn mạnh rằng các hệ thống dựa trên IoT phải dựa trên dữ liệu và tiêu chuẩn mở, bao gồm các giao diện và giao thức, để cho phép đổi mới của bên thứ ba bằng cách giảm thiểu các khóa nhà sản xuất. Dựa trên ý tưởng này, các tác giả thiết kế và phát triển nền tảng GreenIoT ở Thụy Điển để xác định lợi thế của nền tảng mở và dữ liệu mở để phát triển các thành phố thông minh. Tuy nhiên, một số hướng dẫn liên quan đến việc mua sắm cơ sở hạ tầng IoT mở, bao gồm các định dạng dữ liệu phổ biến và giao diện lập trình ứng dụng mở (API), phải được đưa ra.

O. B. Sezer và cộng sự đề xuất một khung tăng cường tích hợp các công nghệ web ngữ nghĩa, dữ liệu lớn và IoT. Các yêu cầu chính cho khung đề xuất đã được phân tích và thiết kế khái niệm của hệ thống IoT có tầm nhìn được đề xuất dựa trên kết quả phân tích. Khung khái niệm bao gồm năm lớp, cụ thể là thu thập dữ liệu, trích xuất-chuyển đổi-tải (ETL), lý luận quy tắc ngữ nghĩa, học tập và hành động. Lớp thu thập dữ liệu, thu thập dữ liệu từ các nguồn khác nhau, có thể được coi là lớp đầu vào của khung. Lớp ETL cung cấp trình điều khiển cảm biến để chuyển đổi dữ liệu nhận được từ các loại cảm biến khác nhau. Lý luận quy tắc ngữ nghĩa hỗ trợ một công cụ lý luận để suy luận từ dữ liệu khung mô tả tài nguyên (RDF) nhận được từ lớp ETL. Lớp học rút ra một số tính năng từ dữ liệu và hình thành các mô hình dựa trên máy học. Lớp hành động cung cấp các hành động được xác định trước cho đầu ra của lớp học.

B. Cheng và cộng sựthiết kế GeeLytics, một nền tảng phân tích cạnh thực hiện xử lý dữ liệu thời gian thực ở các cạnh mạng và trên đám mây. Nền tảng này giải quyết các phân tích địa lý phân phối và độ trễ thấp do số lượng lớn dữ liệu IoT. GeeLytics được thiết kế để hỗ trợ các cấu trúc liên kết xử lý luồng động bằng cách tính đến các đặc điểm hệ thống của các nút cạnh / đám mây không đồng nhất và khối lượng công việc hiện tại của hệ thống.

H. Wang và cộng sự thảo luận về những thách thức và cơ hội do IoT và dữ liệu lớn cho cụm hàng hải. Họ cũng phát triển một khuôn khổ mới để tích hợp IoT công nghiệp với các công nghệ phân tích và dữ liệu lớn. Việc thực hiện khuôn khổ như vậy có thể giúp tăng sản lượng và năng suất cũng như cho phép toàn bộ các cụm tiếp tục đóng vai trò là nhà lãnh đạo trong ngành hàng hải toàn cầu.

Prez và Carrera thực hiện một nghiên cứu toàn diện về đặc tính hiệu năng của API servIoTicy. Họ đặc biệt tập trung vào cơ sở hạ tầng tiên tiến để lưu trữ khối lượng công việc IoT trên đám mây với mục đích cung cấp khả năng xử lý luồng dữ liệu cho nhiều bên thuê, cơ chế truy vấn nâng cao, hỗ trợ đa giao thức và giải pháp phần mềm bằng cách kết hợp nâng cao công nghệ tập trung dữ liệu. Một nghiên cứu khác giải quyết một phần vấn đề lưu trữ dữ liệu lớn bằng cách đề xuất AllJoyn Lambda, một giải pháp phần mềm tích hợp AllJoyn trong kiến ​​trúc Lambda được sử dụng để lưu trữ và phân tích dữ liệu lớn.

A. J. Jara và cộng sự thực hiện một cuộc khảo sát để làm nổi bật các giải pháp và thách thức hiện có đối với dữ liệu lớn được đặt ra bởi các hệ thống vật lý không gian mạng. Nghiên cứu của họ tập trung vào bảo mật đám mây và tích hợp dữ liệu không đồng nhất từ ​​nhiều nguồn. Họ nhấn mạnh sự cần thiết phải phát triển các cơ chế phát hiện dữ liệu tinh vi và để thực hiện xử lý dữ liệu luồng thời gian thực.

Z. Đinh và cộng sự đề xuất một cơ chế cụm cơ sở dữ liệu thống kê chung để phân tích dữ liệu lớn trong mô hình IoT (IOT- StatisticDB). Họ nhập các hàm thống kê trên IOT-StatisticDB thông qua các toán tử thống kê bên trong hệ thống quản lý cơ sở dữ liệu (DBMS). Phân tích thống kê được thực hiện theo kiểu phân tán và song song sử dụng nhiều máy chủ.

C. Vuppalapati và cộng sự kiểm tra vai trò của dữ liệu lớn trong chăm sóc sức khỏe và thấy rằng các cảm biến cơ thể tạo ra một lượng lớn dữ liệu liên quan đến sức khỏe. Hai thách thức được phân tích trong bối cảnh này, đó là tích hợp các điểm dữ liệu giả định này với hồ sơ sức khỏe điện tử (EHR) và trình bày các dữ liệu này cho các bác sĩ trong thời gian thực. Dựa trên những quan sát này, họ đề xuất một khung tích hợp cảm biến cho thấy kiến ​​trúc đám mây có thể mở rộng có thể cung cấp một cách tiếp cận toàn diện cho hệ thống cảm biến EHR. Apache Kafka và Spark được sử dụng để xử lý lượng lớn dữ liệu theo cách thức thời gian thực. Mặc dù hình dung sức khỏe của bệnh nhân trong thời gian thực có thể giúp phát hiện các tình huống khẩn cấp, mô hình này thiếu một giải pháp bảo mật.

A. Ahmad và cộng sự phân tích hành vi của con người bằng cách sử dụng dữ liệu lớn và phân tích trong mô hình IoT xã hội. Họ đề xuất một kiến ​​trúc bao gồm ba lĩnh vực hoạt động. Họ cũng phân tích một hệ sinh thái được tạo ra bởi các thành phố thông minh và dữ liệu lớn. Kỹ thuật quay phim hợp tác có thể được sử dụng trong tương lai để phân tích chính xác hành vi của con người.

D. Arora và cộng sự sử dụng dữ liệu lớn và các kỹ thuật phân tích để phân loại các thiết bị hỗ trợ mạng. Họ cũng phân tích hiệu suất của bốn thuật toán học máy, như k điểm lân cận gần nhất (KNN), NaveBayes (NB), máy vectơ hỗ trợ (SVM) và dự đoán ngẫu nhiên. Kết quả thử nghiệm cho thấy thuật toán NB cho kết quả thấp nhất độ chính xác trong số tất cả các mô hình phân loại, trong khi thuật toán rừng ngẫu nhiên mang lại độ chính xác cao nhất. Trong khi đó, độ chính xác của KNN và SVM liên quan chặt chẽ với thuật toán rừng ngẫu nhiên.

I. L. Yên và cộng sự điều tra tiềm năng của các kỹ thuật khám phá và sáng tác dịch vụ trong việc giải quyết các vấn đề trong thế giới thực dựa trên dữ liệu được tạo thông qua IoT. Họ kiểm tra các công nghệ khác nhau, như phân tích dữ liệu và trí tuệ nhân tạo, có thể được sử dụng trong thế giới thông minh để rút ra các sự kiện tình huống và thực hiện các hành động phù hợp. Họ đề xuất một nền tảng cộng đồng đám đông dựa trên trò chơi để tận dụng trí thông minh của con người để hoàn thành thành công một số nhiệm vụ kiểm soát nhất định. Trong tương lai, các cơ chế chẩn đoán và giám sát chủ động với sự kết hợp khai thác dữ liệu lớn phải được đưa ra để đảm bảo an toàn trong thế giới vật lý thông minh.

R. P. Minch và cộng sự thực hiện một nghiên cứu khám phá về quyền riêng tư của vị trí trong kỷ nguyên của IoT, dữ liệu lớn và phân tích. Họ xác định, phân loại và mô tả các vấn đề riêng tư và tiết lộ các lỗi có thể xảy ra trong bối cảnh dữ liệu lớn và phân tích. Họ đề xuất một khuôn khổ đáng tin cậy để đảm bảo quyền riêng tư trong môi trường nhận biết ngữ cảnh phải được phát triển trong tương lai.

A. Mukherjee và cộng sự đề xuất một khung IoT để thực hiện hiệu quả các công việc phân tích song song dữ liệu. Họ nhằm mục đích xác định một thuật toán phân tích phù hợp có thể đáp ứng các yêu cầu xử lý và phân tích lượng lớn dữ liệu. Phân tích định tính của họ tạo ra kết quả đầy hứa hẹn vì tính hiệu quả cao của các thuật toán phân tích song song trong môi trường IoT. Các nhà nghiên cứu trong tương lai phải giải quyết những vấn đề cản trở việc triển khai mô hình này với sự hiện diện của điện toán sương mù.

A. Mukherjee và cộng sự xuất bản một báo cáo liên quan đến Condor, một khung lưới để thực hiện song song dữ liệu trong mô hình IoT. Kết quả thử nghiệm của họ cho thấy Condor có khả năng mở rộng và sử dụng CPU tốt hơn cho các công việc song song dữ liệu so với kiến ​​trúc dựa trên máy chủ ba tầng truyền thống.

H. R. Arkian và cộng sự đề xuất MIST, một sơ đồ phân tích dữ liệu dựa trên sương mù với phương pháp tối ưu hóa cung cấp tài nguyên hiệu quả về chi phí có thể được sử dụng cho các ứng dụng cảm nhận đám đông IoT. Kế hoạch này nhằm giảm độ trễ cung cấp dịch vụ trong các khung điện toán đám mây truyền thống. Kết quả thử nghiệm cho thấy sơ đồ dựa trên sương mù MIST vượt trội so với tính toán đám mây truyền thống như số lượng ứng dụng yêu cầu dịch vụ thời gian thực tăng lên. Một số giải pháp có thể có của công việc này như sau: (a) thêm mô-đun cảm biến chọn lọc vào lớp sương mù, (b) làm phong phú kiến ​​trúc với khả năng phân tích dữ liệu bảo mật và (c) xem xét tính di động của người tạo dữ liệu và người tiêu dùng dữ liệu trong phần cung cấp tài nguyên.

M. M. Rathore và cộng sự đề xuất một hệ thống xử lý một số vấn đề trong môi trường thành phố thông minh, như cho phép các đối tượng phản ứng với bối cảnh, giảm thiểu chi phí thu thập dữ liệu do thiết bị thông minh tạo ra và thu thập thông tin chi tiết về dữ liệu nếu những dữ liệu này được thu thập và xử lý trong thời gian thực. Hệ thống được đề xuất có kiến ​​trúc bốn tầng, trong đó tầng dưới chịu trách nhiệm tạo và thu thập dữ liệu, tầng 1 trung gian cho phép giao tiếp giữa các cảm biến, rơle, trạm gốc và Internet, tầng 2 trung gian chịu trách nhiệm quản lý dữ liệu và xử lý bằng khung Hadoop và tầng trên cùng chịu trách nhiệm áp dụng các kỹ thuật phân tích dữ liệu và tạo kết quả. Kết quả thực hiện cho thấy hệ thống được đề xuất có khả năng mở rộng và hiệu quả hơn về mặt thông lượng và thời gian xử lý so với các hệ thống hiện tại. Tuy nhiên, hệ thống này thiếu một kỹ thuật ra quyết định thông minh có thể đối phó với dữ liệu lớn trong môi trường IoT.

F. Alam và cộng sự kiểm tra khả năng ứng dụng của tám thuật toán khai thác dữ liệu, bao gồm SVM, KNN, NB, C4.5, C5.0, phân tích phân biệt tuyến tính (LDA), mạng nơ ron nhân tạo (ANN) và ANN (DLANN), tìm hiểu sâu Dữ liệu do IoT tạo ra. Các thuật toán này cũng được so sánh về ma trận nhầm lẫn, độ chính xác phân loại và thời gian thực hiện. Về độ chính xác phân loại, C4.5, C5.0, ANN và DLANN vượt trội so với SVM, KNN, NB và LDA. Tuy nhiên, C4.5, C5.0 và ANN rất giống nhau về độ chính xác phân loại. Trong khi đó, NB và LDA có thời gian thực hiện nhanh nhất, với LDA có thời gian xử lý tốt hơn một chút so với NB. Các tác giả đang có kế hoạch tiến hành một nghiên cứu chi tiết về các bộ dữ liệu IoT lớn hơn và đa dạng trong tương lai. M. H. Berlian và cộng sự giới thiệu một khung giám sát và phân tích lượng lớn dữ liệu được tạo thông qua Internet of Underwater Things (IoUT). Họ sử dụng MapReduce để xử lý các dữ liệu này và thấy rằng MapReduce rút ngắn đáng kể thời gian thực hiện truy vấn so với SQL. Mặc dù có nhiều ưu điểm của khung này, việc thử nghiệm khả năng áp dụng giao thức quản lý tin cậy có thể mở rộng với các ứng dụng IoUT và phát triển kiểm soát sự cho phép dựa trên niềm tin cho các hệ thống IoUT vẫn cần được giải quyết trong tương lai.

D. Mourtzis và cộng sự tiết lộ rằng việc áp dụng IoT trong ngành sản xuất có thể biến đổi các hệ thống truyền thống thành hiện đại. Hơn nữa, việc chuyển đổi như vậy dẫn đến một quy trình sản xuất dữ liệu biến dữ liệu công nghiệp thành dữ liệu lớn trong công nghiệp, vốn bị vô hiệu hóa mà không có sức mạnh phân tích. Việc áp dụng phân tích dữ liệu có thể trao quyền cho các doanh nghiệp để có được các chiến lược dựa trên dữ liệu mới có thể dễ dàng quản lý áp lực cạnh tranh. Họ cũng chứng minh làm thế nào mô hình IoT có thể được thực hiện trong trường hợp đơn giản của một công ty có gần 100 máy.

R. Ramakrish Nam và cộng sự phân tích sự phát triển năng lượng hiện tại ở Ấn Độ và xác định những lợi ích có thể thu được thông qua điện toán đám mây và phân tích. Họ cũng ủng hộ việc sử dụng phân tích có thể cải thiện việc bảo tồn năng lượng, giảm chi phí vận hành và trao quyền cho khách hàng.

1. **Truy cập và phân tích dữ liệu lớn**

Phần này điều tra các nền tảng phân tích và xử lý dữ liệu lớn có thể được sử dụng cho một lượng lớn dữ liệu do IoT tạo ra. Trong IoT, việc xử lý dữ liệu lớn và phân tích dữ liệu có thể được thực hiện gần hơn với nguồn dữ liệu bằng cách sử dụng các dịch vụ của điện toán cạnh di động, đám mây và điện toán sương mù.

1. **Apache Hadoop**

Lần đầu tiên được sử dụng bởi Yahoo! và Facebook, Hadoop là một nền tảng xử lý dữ liệu nguồn mở lưu trữ và xử lý lượng lớn dữ liệu trên một cụm phần cứng hàng hóa. Kiến trúc Hadoop chứa một số thành phần, trong đó quan trọng nhất là Hệ thống tệp phân tán Hadoop (HDFS) và mô hình lập trình MapReduce. HDFS được sử dụng để lưu trữ dữ liệu, trong khi MapReduce được sử dụng để xử lý các dữ liệu này theo cách phân tán. Mặc dù có nhiều ưu điểm, Hadoop thiếu mã hóa ở cấp lưu trữ và mạng, có tính linh hoạt hạn chế, được coi là không phù hợp với các tập dữ liệu nhỏ và có chi phí I / O cao.

1. **1010data**

1010data bao gồm một cơ sở dữ liệu cột và chủ yếu liên quan đến dữ liệu bán cấu trúc, chẳng hạn như dữ liệu IoT. Ngoài khả năng hiển thị dữ liệu, báo cáo và tích hợp dữ liệu, công cụ này cung cấp các dịch vụ phân tích nâng cao, bao gồm tối ưu hóa và phân tích thống kê. 1010data cũng rất hỗ trợ cho cơ sở hạ tầng quy mô lớn. Công cụ này cũng hoạt động theo kiểu tập trung và áp dụng các điều khiển truy cập để tương tác với các hệ thống back-end. 1010data có thể đáp ứng nhu cầu của khách hàng thông qua các khả năng phân tích tiên tiến của nó. Tuy nhiên, 1010data được coi là không hiệu quả về mặt khai thác, chuyển đổi và tải dữ liệu.

1. **Cloudera Data Hub**

Cloudera đã giới thiệu Enterprise Data Hub, một khung dựa trên Hadoop để xử lý và phân tích dữ liệu IoT lớn có thể được sử dụng như một điểm trung tâm trong việc quản lý số lượng lớn dữ liệu IoT từ các doanh nghiệp. Để đạt được độ tin cậy, hiệu suất cao, bảo mật và kiểm soát truy cập dữ liệu, Cloudera Data Hub kết hợp Cloudera Manager, Navigator và các thành phần sao lưu và phục hồi của nó. Tuy nhiên, công cụ này không có hệ thống phần cứng và phần mềm riêng và chỉ dựa vào bên thứ ba khi xác định các vấn đề bảo mật và quyền riêng tư nghiêm trọng.

1. **SAP-Hana**

SAP-Hana là một nền tảng trong bộ nhớ để thực hiện phân tích dữ liệu IoT lớn và giải quyết các nhu cầu giao dịch. SAP hỗ trợ các giải pháp phân tán khác nhau để chứa dữ liệu phi cấu trúc lớn. Hana truy cập dữ liệu lớn thông qua Hive, trong khi SAP sử dụng Sybase IQ để cung cấp DBMS cột. Hana cũng có thư viện phân tích tích hợp để chứa, xử lý không gian và hỗ trợ thư viện phân tích văn bản và ngôn ngữ R. Ngoài độ trễ thấp, SAP-Hana còn có thể phân tích cả dữ liệu văn bản và dữ liệu phi cấu trúc. Tuy nhiên, trong công cụ này, tất cả dữ liệu trong một hàng phải được đọc mặc dù chỉ có dữ liệu từ một vài cột được yêu cầu truy cập. Hơn nữa, khả năng của SAP-Hana không đủ mạnh so với các giải pháp khác.

1. **HP-HAVEn**

HP đã giới thiệu bảo mật Hadoop Autonomy Vertica Enterprise (Lawrn), một kiến ​​trúc nền tảng dữ liệu IoT lớn mới cho một số lượng lớn các hệ thống HP có thể được sử dụng với bất kỳ số lượng ứng dụng nào. HP cung cấp cấu hình phần cứng tham chiếu cho các nhà phân phối chính của phần mềm Hadoop. Phần mềm Autolty ID IDOL cung cấp dịch vụ tìm kiếm và khám phá dữ liệu phi cấu trúc. Vertica là một DBMS phân tích cho cơ sở dữ liệu cột xử lý song song ồ ạt nhằm tăng tốc độ phân tích các bộ dữ liệu có cấu trúc lớn. HP Lawrn hiện đang hợp tác với một số công ty để bổ sung cho kho dữ liệu doanh nghiệp cũ. HP cũng giới thiệu một loại Flex-Zone để tạo điều kiện cho việc khám phá các bộ dữ liệu lớn trước khi xác định sơ đồ cơ sở dữ liệu. Hạn chế duy nhất của HP-Lawrn là sự gia tăng số lượng khách thuê tạo ra một danh mục cơ sở dữ liệu lớn trong đó các quy trình giữ và giải phóng khóa trong tất cả các hoạt động đều được giảm tốc.

1. **Hortonworks**

Hortonworks tập trung vào việc xây dựng một nền tảng quản lý và phân tích dữ liệu IoT lớn dựa trên Hadoop. Nền tảng dữ liệu Hortonworks (HDP) có phân phối phần mềm nguồn mở miễn phí và tập trung vào việc cải thiện Hive. Tuy nhiên, với plugin HDP của mình, Hortonworks không thể giảm số lượng nhóm nút hoặc máy chủ cho mỗi nhóm nút trong cụm được tạo.

1. **Bộ dữ liệu lớn Pivotal**

Bộ dữ liệu lớn Pivotal (Pivotal BDS), thường được triển khai trong đám mây công cộng, bao gồm ba giải pháp, đó là Pivotal HDB, Pivotal Greenplum và Pivotal GemFire, tất cả đều được phân phối theo một giấy phép duy nhất. Pivotal là một cơ sở dữ liệu phân tích kết hợp hiệu suất phân tích dựa trên xử lý song song (MPP) với tuân thủ SQL ANSI mạnh mẽ và giúp thực hiện phân tích dự đoán trên dữ liệu được lưu trữ trong HDFS bằng cú pháp SQL và các công cụ liên quan khác. Pivotal Greenplum là một cơ sở dữ liệu phân tích MPP mã nguồn mở, được sử dụng để thực hiện phân tích nhanh trên lượng dữ liệu khổng lồ và cung cấp hiệu suất truy vấn cao trên khối lượng dữ liệu ở quy mô petabyte. Pivotal GemFire ​​là một mạng dữ liệu inmemory được thiết kế để hỗ trợ khối lượng lớn các ứng dụng giao dịch và vận hành. Mặc dù có rất nhiều lợi ích, Pivotal BDS vẫn còn ở giai đoạn sơ khai và việc áp dụng rộng rãi của nó bị sa lầy bởi nhiều vấn đề chưa được giải quyết.

1. **Infobright**

Một công cụ được thiết kế đặc biệt để giải quyết vấn đề quản lý dữ liệu và phân tích, Infobright có thể phân tích tới 50 terabyte dữ liệu. Với tỷ lệ nén và bỏ qua dữ liệu cao, Infobright được coi là phù hợp với dữ liệu do máy tạo, chẳng hạn như dữ liệu IoT. Infobright chủ yếu hoạt động với Hadoop hoặc kho dữ liệu quy mô cao. Công nghệ bỏ qua dữ liệu và thiết kế cột của công cụ này đảm bảo rằng chỉ các dữ liệu liên quan sẽ được sử dụng trong mỗi truy vấn. Những dữ liệu này cũng được lập chỉ mục tự động mà không cần bất kỳ phân vùng và điều chỉnh. Tuy nhiên, tất cả các truy vấn không thể được trả lời tối ưu bằng trình tối ưu hóa Infobright.

1. **MAPR**

MapR [43] hỗ trợ dữ liệu và phân tích lớn cũng như áp dụng một số thành phần của Hadoop để cải thiện hiệu suất của nó (ví dụ: thay thế HDFS bằng hệ thống tệp mạng giống như NFS để đạt được bảo mật và tính sẵn sàng cao). MapR cũng có phương pháp phục hồi hệ thống riêng.

Gần đây, MapR đã thêm các tùy chọn xử lý luồng và tìm kiếm LucidWorks vào Hadoop để tăng cường khả năng dự đoán và cho phép xử lý nhanh. Tuy nhiên, MapR có độ phức tạp cao hơn so với Hadoop.

1. **Yêu cầu**

Các yêu cầu của dữ liệu lớn và phân tích trong IoT đã tăng theo cấp số nhân trong những năm qua và hứa hẹn những cải tiến đáng kể trong quá trình ra quyết định. Do đó, nhu cầu thích ứng phân tích dữ liệu với dữ liệu lớn trong IoT cũng tăng lên, do đó thay đổi cách thu thập, lưu trữ và phân tích dữ liệu. Dữ liệu lớn và phân tích có tiềm năng lớn để trích xuất thông tin có ý nghĩa từ dữ liệu được tạo ra bởi các thiết bị cảm biến. Các yêu cầu chung đối với dữ liệu lớn và IoT xác định các thông số kỹ thuật chức năng và không chức năng cho phân tích dữ liệu. Phần này trình bày các yêu cầu chính cho dữ liệu lớn và phân tích trong môi trường IoT. Những yêu cầu này đóng vai trò quan trọng trong việc cải thiện dịch vụ IoT thông qua phân tích.

1. **Kết nối**

Mô hình IoT đang dần dẫn đến sự kết nối phổ biến của các đối tượng được trang bị cảm biến thông minh trong một môi trường thông minh. Một trong những yêu cầu chính của IoT là cung cấp khả năng kết nối đáng tin cậy cho dữ liệu lớn và phân tích để tạo điều kiện cho việc kết hợp và tích hợp khối lượng lớn dữ liệu cảm biến do máy tạo ra. Do đó, nhiều thiết bị xung quanh chúng ta có tiềm năng lớn để được kết nối với cơ sở hạ tầng điện toán hiệu năng cao để cung cấp các dịch vụ IoT. Ngoài ra, với sự hiện diện ngày càng tăng của truy cập Internet không dây WiFi và 4G-LTE, sự phát triển hướng tới các mạng thông tin và truyền thông phổ biến đã được thể hiện rõ. Tuy nhiên, một kết nối liền mạch giữa các đối tượng khác nhau trong các thành phố thông minh, như IoT, điện toán đám mây, dữ liệu lớn và phân tích, phải được thiết lập trước khi đưa trí thông minh vào môi trường của chúng ta.

1. **Lưu trữ**

Sự tăng trưởng nhanh chóng liên tục của một số lượng lớn các đối tượng hỗ trợ IoT đã dẫn đến việc lưu trữ một lượng lớn dữ liệu không đồng nhất trong phần cứng hàng hóa chi phí thấp trên cơ sở thời gian thực. Các yêu cầu chính của lưu trữ dữ liệu lớn trong IoT bao gồm xử lý một lượng rất lớn dữ liệu phi cấu trúc và cung cấp độ trễ thấp cho các phân tích. Hơn nữa, các ứng dụng của công nghệ dữ liệu lớn cho IoT có thể cho phép lưu trữ và xử lý dữ liệu hiệu quả để tạo ra thông tin có thể tăng cường các dịch vụ thành phố thông minh khác nhau. Phổ của các nguồn dữ liệu IoT bao gồm dữ liệu cảm biến, điện thoại thông minh và phương tiện truyền thông xã hội được mô hình hóa theo các cách khác nhau và sử dụng các giao thức và giao diện truyền thông khác nhau. Hầu hết các dịch vụ IoT đều dựa trên các giao thức truyền thông M2M, yêu cầu xử lý một số lượng lớn luồng và trực tiếp hưởng lợi từ khả năng lưu trữ phân tán rộng rãi của cơ sở hạ tầng điện toán đám mây.

1. **Chất lượng dịch vụ**

Quản lý tài nguyên của cảm biến IoT và thiết bị di động là yêu cầu chính đối với chất lượng dịch vụ (QoS) để phân tích hiệu quả một lượng dữ liệu khổng lồ. Mặc dù nhiều nghiên cứu đã cố gắng đáp ứng yêu cầu QoS, làm thế nào để thống nhất và tích hợp kiến ​​trúc QoS vào IoT để hỗ trợ dữ liệu lớn và phân tích đảm bảo nghiên cứu thêm. QoS được cung cấp bởi mạng IoT phải đáng tin cậy và phải đảm bảo truyền dữ liệu di động và hiệu quả từ các nguồn đó, nơi tạo ra dữ liệu lớn. Hỗ trợ QoS trong mạng này cực kỳ quan trọng đối với dữ liệu và phân tích lớn. Tuy nhiên, để tạo ra một mạng đáng tin cậy, nhiều công nghệ mạng mới nổi phải được đưa vào IoT để cho phép chuyển sự kiện theo thời gian thực và cải thiện khả năng xử lý dữ liệu lớn.

1. **Phân tích thời gian thực**

Phân tích phát trực tuyến đã nhanh chóng nổi lên như một sáng kiến ​​IoT quan trọng cho các quá trình ra quyết định kịp thời [48]. Một trong những tính năng nổi bật nhất của IoT là giao tiếp thông tin theo thời gian thực hoặc gần thời gian liên quan đến liên kết với nhau

nhiều thứ". Dữ liệu lớn và phân tích trong IoT yêu cầu phát trực tiếp các sự kiện và lưu trữ dữ liệu phát trực tuyến trong cơ sở dữ liệu hoạt động. Do phần lớn các dữ liệu phi cấu trúc này được truyền trực tiếp từ các thứ được kích hoạt trên web, các triển khai dữ liệu lớn phải thực hiện phân tích với các truy vấn thời gian thực để giúp các tổ chức hiểu nhanh, đưa ra quyết định và tương tác với mọi người và các thiết bị khác trong thời gian thực [15].

1. **Điểm chuẩn**

Dữ liệu lớn và phân tích đã thu hút nhiều sự chú ý từ các học viện và các tổ chức khác nhau, và nhiều tổ chức cũng đã bắt đầu theo đuổi các doanh nghiệp IoT. Tuy nhiên, các tổ chức này phải đối mặt với một số thách thức trong việc lưu trữ và phân tích lượng lớn dữ liệu được thu thập thông qua các cảm biến trong môi trường IoT. Giải quyết những vấn đề này đòi hỏi một sự hiểu biết sâu sắc có thể đạt được bằng cách sử dụng một nền tảng phân tích và dữ liệu lớn. Điểm chuẩn đóng vai trò quan trọng trong bối cảnh này bằng cách cung cấp cho các tổ chức một cách để đánh giá chất lượng của các giải pháp phân tích và dữ liệu lớn. Một điểm chuẩn hệ thống tuyệt vời cũng có thể cung cấp các so sánh đơn giản và dễ hiểu về các giải pháp khác nhau.

1. **Phân loại**

Hình 4 cho thấy sự phân loại theo chủ đề của các giải pháp dữ liệu lớn và phân tích được thiết kế cho các hệ thống IoT. Các giải pháp này được phân loại dựa trên các thuộc tính sau: a) nguồn dữ liệu lớn, b) thành phần hệ thống, c) công nghệ cho phép dữ liệu lớn, d) yếu tố chức năng và e) loại phân tích.

1. **Nguồn dữ liệu lớn**

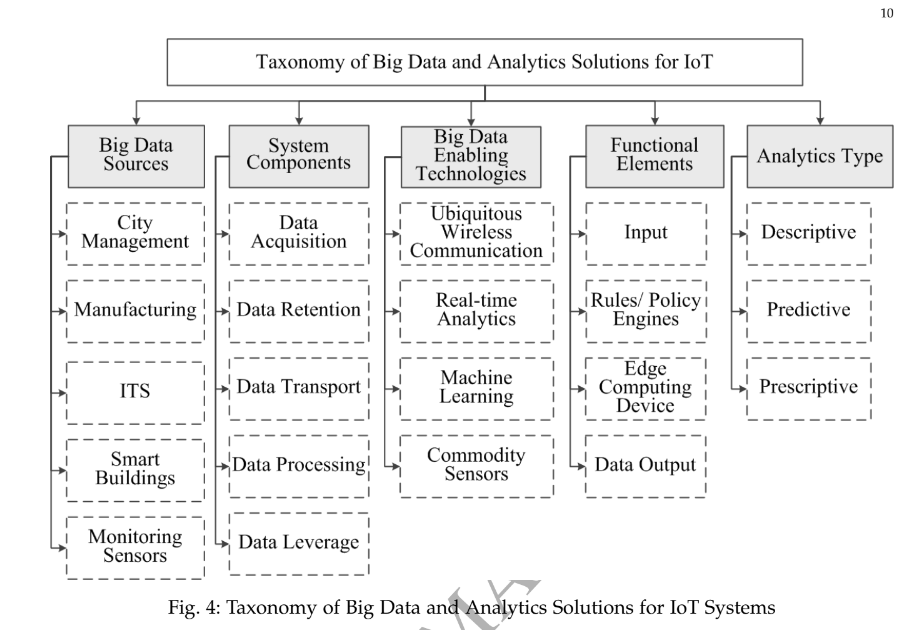
Dữ liệu lớn được tạo ra bởi cơ sở hạ tầng được triển khai để chạy các ứng dụng IoT khác nhau, bao gồm quản lý thành phố, sản xuất, hệ thống giao thông thông minh (ITS), tòa nhà thông minh và cảm biến giám sát.

Ban quản lý thành phố sử dụng các camera, cảm biến và thiết bị truyền động được kết nối để làm cho cuộc sống của người dân được an toàn và thuận tiện. Tuy nhiên, các thiết bị này tạo ra một lượng lớn dữ liệu phải được quản lý và phân tích trong thời gian thực để có được thông tin chi tiết có liên quan. Tương tự, ngành công nghiệp sản xuất đã triển khai các thiết bị IoT liên tục tạo ra một lượng dữ liệu khổng lồ để tối đa hóa năng suất và hiệu quả của hoạt động. Để có được cái nhìn sâu sắc từ những dữ liệu này, các giải pháp dữ liệu và phân tích lớn đã được sử dụng trong việc thiết kế và thử nghiệm các sản phẩm mới, tối ưu hóa dịch vụ và tiếp thị, giảm thiểu lỗi và cải thiện năng suất.

Cùng với dữ liệu và phân tích lớn, sự phổ biến của các cảm biến, các công nghệ xe được kết nối và IoT đã dẫn đến việc tạo ra các hệ thống giao thông thông minh, do đó làm tăng đáng kể lượng thời gian thực dữ liệu lớn phải được truyền đạt, tổng hợp, phân tích và quản lý. ITS có thể tận dụng dữ liệu lớn và phân tích để tăng cường khả năng ra quyết định của người dùng.

Việc sử dụng các giải pháp dữ liệu lớn trong tòa nhà thông minh có cùng mục tiêu như trong các lĩnh vực ứng dụng khác. Thông tin liên quan được trích xuất từ ​​một loạt các dữ liệu hiện có và sau đó được cung cấp cho những người ra quyết định để quản lý dịch vụ và cho người dùng của tòa nhà.

Dữ liệu lớn trong môi trường IoT thường được sử dụng để thu thập và lưu trữ dữ liệu cảm biến giám sát, thực hiện phân tích dữ liệu, đưa ra dự báo và tạo cảnh báo nếu phát hiện sai lệch bất thường.



1. **Các thành phần hệ thống**

Các giải pháp phân tích và dữ liệu lớn thường bao gồm năm thành phần hệ thống, cụ thể là thu thập dữ liệu, lưu trữ dữ liệu, vận chuyển dữ liệu, xử lý dữ liệu và sử dụng dữ liệu.

Thu thập dữ liệu lớn liên quan đến việc thu thập, lọc và làm sạch dữ liệu trước khi chúng được chuyển vào kho dữ liệu. Thành phần này thường được chi phối bởi bốn thuộc tính, cụ thể là khối lượng, giống, vận tốc và giá trị. Lưu giữ dữ liệu lớn liên quan đến các chính sách còn tồn tại và yêu cầu ban quản lý đáp ứng các yêu cầu lưu trữ dữ liệu lớn. Các chính sách lưu giữ dữ liệu lớn khác nhau liên quan đến quyền riêng tư và các vấn đề pháp lý đối với kinh tế để xác định các quy tắc lưu trữ, thời gian lưu giữ, định dạng dữ liệu và phương thức mã hóa. Dữ liệu lớn phải được vận chuyển qua các trang dữ liệu khác nhau để đảm bảo cân bằng tải, liên tục kinh doanh và nhân rộng. Dữ liệu lớn là một thuật ngữ được sử dụng cho các bộ dữ liệu lớn và phức tạp không thể xử lý bằng phần mềm truyền thống. Những thách thức chính liên quan đến xử lý dữ liệu lớn có liên quan đến việc nắm bắt, lưu trữ, phân tích, tìm kiếm, cập nhật, trực quan hóa và quyền riêng tư. Đòn bẩy dữ liệu lớn liên quan đến việc đảm bảo làm thế nào một doanh nghiệp có thể gặt hái lợi ích từ dữ liệu của họ để tăng doanh thu của họ.

1. **Công nghệ kích hoạt dữ liệu lớn**

Dữ liệu lớn cho phép các công nghệ trong bối cảnh IoT có liên quan đến tính năng không dây phổ biến, phân tích thời gian thực, học máy và các yếu tố thu thập dữ liệu, như cảm biến hàng hóa và hệ thống nhúng.

Các công nghệ truyền thông không dây phổ biến có mặt khắp nơi được sử dụng để vận chuyển dữ liệu lớn trong IoT bao gồm IEEE 802.15.4, IEEE 802.11, IEEE 802.15.1 và IEEE 802.16.

Phân tích thời gian thực làm cho dữ liệu lớn được tạo bởi các thiết bị IoT sẵn sàng sử dụng khi chúng vào hệ thống. Thời gian thực có thể được định nghĩa là mức độ phản ứng của máy tính là tức thời hoặc gần như tức thời.

Không giống như các công cụ phân tích truyền thống, học máy có thể khai thác những hiểu biết ẩn giấu trong dữ liệu lớn và trích xuất các giá trị từ các nguồn dữ liệu lớn với sự tương tác tối thiểu của con người. Học máy rất phù hợp trong bối cảnh IoT vì các nguồn dữ liệu khác nhau và số lượng lớn và nhiều loại dữ liệu liên quan.

Dữ liệu lớn trong IoT được thu thập bằng cách sử dụng một số cảm biến và cơ cấu chấp hành. Các công nghệ cảm biến này có vai trò chính trong việc thu thập và truyền dữ liệu đến các tài nguyên cạnh gần đó để xử lý thêm.

1. **Các yếu tố chính**

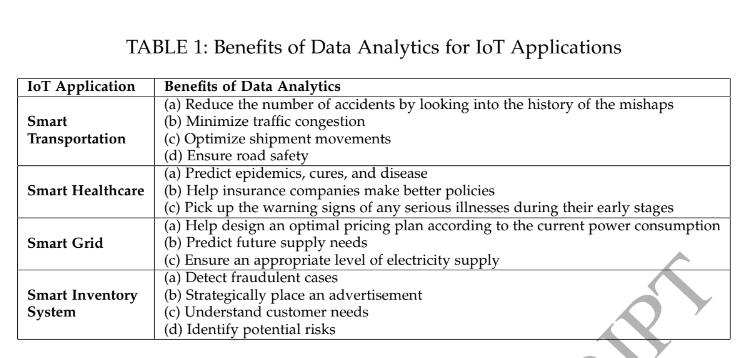
Các giải pháp phân tích và dữ liệu lớn cho IoT bao gồm bốn yếu tố chính là đầu vào, quy tắc hoặc công cụ chính sách, thiết bị tính toán cạnh và đầu ra dữ liệu. Dữ liệu thô được thu thập từ các tài nguyên khác nhau và được chuyển đến các hệ thống phân tích cạnh. Các hệ thống này dựa trên một quy tắc / công cụ chính sách xác định và áp dụng các quy tắc cho dữ liệu đầu vào để có được thông tin chi tiết. Thiết bị tính toán cạnh là một nhân tố quan trọng khác trong hoạt động của hệ thống phân tích. Xử lý dữ liệu được tạo bởi các thiết bị IoT trên các thiết bị cạnh có thể mang lại một số lợi thế, như độ trễ thấp, tiêu thụ băng thông tối thiểu, tính toàn vẹn dữ liệu, bảo mật và chi phí thấp. Những dữ liệu này cũng được cung cấp cho người tiêu dùng trong thời gian thực.

1. **Loại phân tích**

Phân tích có thể được chia thành ba loại, cụ thể là phân tích mô tả, phân tích dự đoán và phân tích phối cảnh. Các phân tích mô tả, định nghĩa về những gì đã xảy ra hoặc những gì đang xảy ra, đã giúp tìm ra những cơ hội và thách thức kinh doanh mới. Các phân tích dự đoán, định nghĩa về những gì sẽ xảy ra và tại sao nó sẽ xảy ra, thì được kích hoạt bằng cách sử dụng các công nghệ khác nhau, như khai thác văn bản / web / dữ liệu, để dự đoán chính xác các điều kiện và trạng thái trong tương lai. Các phân tích theo quy định, trong đó xác định về những gì tôi nên làm và tại sao tôi nên làm điều đó, bởi vì sử dụng các hệ thống mô phỏng, chuyên môn và hỗ trợ quyết định để điều tra các lựa chọn khác nhau và đưa ra đề xuất cho những người ra quyết định.

1. **VAI TRÒ CỦA PHÂN TÍCH DỮ LIỆU TRONG ỨNG DỤNG IOT**

Các công nghệ dữ liệu lớn có thể cung cấp dịch vụ lưu trữ và xử lý dữ liệu trong môi trường IoT, trong khi phân tích dữ liệu cho phép người kinh doanh đưa ra quyết định tốt hơn. Các ứng dụng IoT là nguồn chính của dữ liệu lớn. Phần này giải thích vai trò của dữ liệu lớn và phân tích trong các ứng dụng IoT khác nhau, bao gồm lưới thông minh, chăm sóc sức khỏe thông minh, vận chuyển thông minh và thông minh trong các hệ thống thông hơi. Bảng 1 tóm tắt lợi ích của dữ liệu lớn và phân tích trong các ứng dụng IoT.



1. **Giao thông thông minh**

Tìm kiếm thông tin có giá trị đã trở thành mối quan tâm chính trong thời đại công nghệ hiện đại nơi các phương tiện được kết nối với Internet và tạo ra một lượng lớn dữ liệu. Phân tích dữ liệu có thể giúp các cơ quan quản lý vận tải tìm ra lịch sử của các tai nạn trên đường (ví dụ: trong trường hợp nào đã xảy ra tai nạn và các tài xế lái xe trong lúc xảy ra tai nạn), giảm thiểu số vụ tai nạn đường bộ, xác định thời điểm xảy ra tải lưu lượng đạt đến đỉnh điểm và chuẩn bị một kế hoạch tuyến tối ưu có thể giúp giảm thiểu tắc nghẽn giao thông.

Việc phân tích dữ liệu vận chuyển thông minh có thể gián tiếp tối ưu hóa các chuyển động của lô hàng, cải thiện an toàn đường bộ và nâng cao trải nghiệm người dùng từ đầu đến cuối về thời gian giao hàng.

1. **Chăm sóc sức khỏe thông minh**

Trong vài năm qua, lượng dữ liệu khổng lồ đã được tạo ra trong lĩnh vực chăm sóc sức khỏe. Tuy nhiên, sự gia tăng nhanh chóng trong sản xuất dữ liệu đã tạo ra những thách thức trong việc trích xuất thông tin có giá trị từ các bộ dữ liệu chăm sóc sức khỏe lớn có thể giúp dự đoán dịch bệnh và tìm ra cách chữa trị cho các bệnh khác nhau. Phân tích dữ liệu có thể giúp các chuyên gia chăm sóc sức khỏe phân tích một lượng lớn dữ liệu bệnh nhân và tìm hiểu lịch sử của bệnh (trong trường hợp bác sĩ gia đình). Các công ty bảo hiểm cũng có thể sử dụng phân tích dữ liệu khi đưa ra chính sách. Các chuyên gia chăm sóc sức khỏe cũng có thể phát hiện các bệnh nghiêm trọng ở giai đoạn đầu và sau đó ngăn ngừa mất mạng.

1. **Lưới thông minh**

Các lưới thông minh nhanh chóng tạo ra dữ liệu và việc tìm kiếm thông tin hữu ích từ các dữ liệu này đã trở nên cấp thiết. Trong môi trường lưới điện thông minh, một lượng lớn dữ liệu được thu thập từ nhiều nguồn khác nhau, chẳng hạn như thói quen sử dụng năng lượng của người dùng, dữ liệu đo lường phasor để nhận thức tình huống và dữ liệu tiêu thụ năng lượng được đo bằng đồng hồ thông minh rộng rãi, để nêu tên một số. Phân tích thích hợp có thể giúp người ra quyết định đo mức cung cấp điện phù hợp mà họ phải cung cấp cho khách hàng. Phân tích cũng có thể giúp người kinh doanh dự đoán nhu cầu điện trong tương lai gần. Các mục tiêu chiến lược của các tổ chức cụ thể cũng có thể được đáp ứng thông qua các phân tích thích hợp (ví dụ: các kế hoạch giá phù hợp với mô hình cung, cầu và sản xuất).

1. **Hệ thống kiểm kê thông minh**

Tìm kiếm thông tin hữu ích từ một lượng lớn dữ liệu hệ thống hàng tồn kho có thể giúp chủ doanh nghiệp tạo ra nhiều lợi nhuận hơn. Việc phân tích các bộ dữ liệu do hệ thống kiểm kê tạo ra có thể giúp người ta có được kiến ​​thức về xu hướng thị trường. Khuyến nghị sản phẩm có thể được tạo ra sau khi phân tích các biến thể theo mùa. Việc phân tích dữ liệu hàng tồn kho cũng có thể giúp phát hiện các trường hợp gian lận. Phân tích có thể hỗ trợ các nhà quảng cáo trong chiến lược đặt quảng cáo của họ. Phân tích dự đoán có thể giúp mọi người đưa ra quyết định có giá trị và hiểu thêm về khách hàng và sản phẩm của họ. Phân tích dữ liệu cũng có thể giúp các công ty xác định các rủi ro và cơ hội tiềm năng của họ.

1. **CƠ HỘI**

Môi trường IoT hiện tại cung cấp các cơ hội sau cho dữ liệu và phân tích lớn hiệu quả.

1. **Ra quyết định**

Sự phổ biến của các thiết bị IoT, điện thoại thông minh và phương tiện truyền thông xã hội mang đến cho những người ra quyết định cơ hội trích xuất thông tin có giá trị về người dùng của họ, để dự đoán xu hướng trong tương lai và phát hiện gian lận. Dữ liệu lớn có thể tạo ra giá trị đáng kể bằng cách làm cho thông tin minh bạch và có thể sử dụng được cho các tổ chức, do đó giúp họ phơi bày sự thay đổi và tăng hiệu suất của họ. Phần lớn dữ liệu được tạo thông qua IoT và các công cụ phân tích khác nhau tạo ra một số lượng lớn cơ hội cho các tổ chức. Những công cụ này tận dụng mô hình dự báo, phân loại và kỹ thuật phân cụm để cung cấp các giải pháp khai thác dữ liệu khác nhau. Khai thác IoT cũng có thể cải thiện thói quen ra quyết định của các cá nhân sử dụng dữ liệu lớn.

1. **Cải thiện hiệu quả**

Nhu cầu xử lý và lưu trữ dữ liệu của các ứng dụng phân tích nâng cao đã hạn chế việc áp dụng chúng trong nhiều lĩnh vực. Tuy nhiên, những rào cản như vậy đang bắt đầu giảm vì IoT. Các công nghệ dữ liệu lớn, như Hadoop và các công cụ khai thác dựa trên đám mây, mang lại lợi thế đáng kể về mặt giảm chi phí so với các kỹ thuật khai thác truyền thống. Hơn nữa, các kỹ thuật phân tích truyền thống đòi hỏi dữ liệu phải ở một định dạng nhất định, rất khó đạt được khi sử dụng dữ liệu IoT. Tuy nhiên, sử dụng các công nghệ dữ liệu lớn hiện tại xây dựng xung quanh phần cứng cộng đồng chi phí thấp có thể giúp cải thiện khả năng phân tích và giảm chi phí xử lý.

1. **Độc lập với Silo dữ liệu**

Sự ra đời của IoT và cho phép các công nghệ như điện toán đám mây đã cho phép loại bỏ các silo dữ liệu trong các lĩnh vực khác nhau. Thông thường, mỗi loại dữ liệu được coi là hữu ích

chỉ cho miền của nó, nhưng dữ liệu tên miền chéo đã nổi lên như là giải pháp hiệu quả cho các vấn đề khác nhau. Các loại dữ liệu khác nhau, chẳng hạn như dữ liệu thời gian chạy, siêu dữ liệu thiết bị, dữ liệu thương mại, dữ liệu bán lẻ và dữ liệu doanh nghiệp, hiện có thể được sử dụng do các công nghệ cho phép khác nhau bổ sung cho IoT, bao gồm dữ liệu lớn, đám mây, web ngữ nghĩa và công nghệ lưu trữ dữ liệu .

1. **Ứng dụng giá trị gia tăng**

Học sâu, học máy và trí tuệ nhân tạo là những công nghệ chính cung cấp các ứng dụng giá trị gia tăng sử dụng IoT và dữ liệu lớn. Trước khi xuất hiện IoT và điện toán đám mây, một lượng lớn dữ liệu và sức mạnh tính toán không có sẵn cho một số ứng dụng nhất định, do đó ngăn chúng sử dụng các công nghệ như vậy. Các nền tảng phân tích dữ liệu khác nhau, nền tảng thông minh kinh doanh, ứng dụng trực quan và ứng dụng phân tích gần đây đã xuất hiện và giúp các ngành công nghiệp và tổ chức chuyển đổi hoạt động, cải thiện năng suất và chẩn đoán và tăng tính linh hoạt. Mức độ chi tiết như vậy là không thể trước khi IoT xuất hiện.

1. **THÁCH THỨC NGHIÊN CỨU MỞ**

Các hệ thống IoT có tiềm năng để giải quyết nhiều vấn đề, nhưng nhiều thách thức vẫn chưa được giải quyết. Các giải pháp cho một số thách thức này vẫn chưa được cung cấp bởi chính các giải pháp phân tích và dữ liệu lớn, trong khi các giải pháp khác đòi hỏi nỗ lực tập trung từ cộng đồng IoT, nhà cung cấp phần cứng và nền tảng, chính phủ và nhà hoạch định chính sách.

1. **Khai thác tính hữu dụng tạm thời của IoT**

Dữ liệu IoT có tác động sâu sắc đến thế giới số hóa. Tuy nhiên, những dữ liệu này có khía cạnh tạm thời có thể hữu ích trong việc đưa ra quyết định thời gian thực, cải thiện chất lượng và cung cấp trải nghiệm người dùng tuyệt vời. Ví dụ, một tổ chức tiêu dùng có thể kết hợp dữ liệu tiêu dùng có sẵn với dữ liệu chiếm dụng bãi đậu xe hàng ngày để cung cấp các ưu đãi cho khách hàng của họ hoặc quản lý hàng tồn kho của họ một cách chủ động hàng ngày hoặc theo mùa. Trong các giải pháp IoT điển hình, những hiểu biết sâu sắc từ dữ liệu IoT thường tốn thời gian hoặc không được đưa vào sử dụng ngay lập tức. Xu hướng này thay đổi thành một xu hướng chủ động để tạo ra các mối tương quan, rút ​​ra những hiểu biết sâu sắc và tìm ra các mô hình theo mùa, mới nổi và giảm dần bằng cách sử dụng dữ liệu IoT. Trong nhiều ứng dụng công nghiệp quan trọng, những tương quan, hiểu biết và mô hình này có thể giúp tăng hiệu quả hoạt động và đạt được sự kiểm soát hiệu quả trong thời gian thực. Do đó, chúng tôi phải triển khai các giải pháp có thể xử lý dữ liệu ở cấp thiết bị hoặc cổng nơi dữ liệu IoT từ thiết bị, cảm biến và quy trình ban đầu được nhận. Khai thác dữ liệu chú thích ngữ nghĩa hoặc sử dụng công cụ quy tắc để xử lý thông tin cục bộ là những con đường tiềm năng để khám phá trong nghiên cứu trong tương lai. Áp dụng ngữ nghĩa đặc biệt hữu ích vì khả năng cung cấp các khái niệm trừu tượng cần thiết, trong khi dữ liệu chú thích vẫn giữ được ngữ nghĩa của chúng khi được đẩy lên nền tảng IoT / đám mây để phân tích.

1. **Thêm ngữ nghĩa vào dữ liệu IoT**

Tính hữu ích của bất kỳ loại dữ liệu nào cũng có thể được tăng cường bằng cách thêm siêu dữ liệu vào ngữ cảnh và ý nghĩa của chúng. Thực tiễn này đặc biệt quan trọng trong IoT bằng cách giúp người dùng xử lý và sử dụng dữ liệu IoT không đồng nhất ở thiết bị, cổng và cấp độ đám mây với các phạm vi và mức độ chi tiết khác nhau. Một lựa chọn là dựa trên các giải pháp về bản thể học của họ, đó là một đại diện chính thức của các khái niệm và các mối quan hệ giữa các khái niệm này. Do đó, bản thể học có thể được sử dụng để tạo ra các từ vựng của siêu dữ liệu để chú thích dữ liệu IoT tại nguồn hoặc gần nguồn. Cho rằng các bản thể học rất dễ chia sẻ và liên kết, chúng có thể cung cấp ngữ cảnh và ý nghĩa đúng của dữ liệu IoT theo cách mở. Bản thể cũng hữu ích cho việc tích hợp dữ liệu IoT từ nhiều miền. Mặc dù một số nỗ lực đã được thực hiện để tạo ra các bản thể chung và miền cụ thể, nhưng cần có nhiều nỗ lực hơn trong một số lĩnh vực, đặc biệt là trong thế giới công nghiệp, để tạo ra các bản thể cụ thể để liên kết và chia sẻ dữ liệu

lĩnh vực và doanh nghiệp khác nhau. Một tùy chọn khác là sử dụng các từ vựng phi bản thể học, chẳng hạn như dự án Haystack, tập trung vào việc xác định các thẻ siêu dữ liệu để chú thích trong miền tự động hóa tòa nhà.

Tuy nhiên, tùy chọn này thiếu tích hợp với các từ vựng tương tự khác. Người ta cũng có thể gặp một số từ vựng rời rạc trừ khi chúng phát triển thành các bản thể có thể được liên kết và chia sẻ trên các miền. Một tùy chọn khác là sử dụng các tiêu chuẩn mở, chẳng hạn như một tiêu chuẩn từ tập đoàn Hypercat sử dụng định dạng danh mục tiêu chuẩn để mã hóa siêu dữ liệu dưới dạng bộ ba RDF và liên kết chúng với nhau bằng cách sử dụng URL. Tuy nhiên, những nỗ lực như vậy vẫn chưa được mở rộng đến cấp độ toàn cầu.

1. **Các vấn đề đa dạng**

Mô hình IoT có các giao thức, tiêu chuẩn và nền tảng không đồng nhất. Thế giới công nghiệp cũng phải đối mặt với các vấn đề tích hợp CNTT và OT thể hiện sự phân mảnh công nghệ. Các giao thức hiện tại có một số sáng kiến, bao gồm CoAP, MQTT, XMPP, DDS, STOMP, HTTP và AMQP. Mặc dù mô hình IoT không có giao thức phổ quát, nhiều giao thức có thể cùng tồn tại do các yêu cầu khác nhau và mục đích sử dụng của chúng. Do đó, các hệ thống IoT có thể không thể hỗ trợ nhiều giao thức theo cách mở rộng. Các giải pháp cổng thông minh, như được đề xuất trong [69], phải cung cấp khả năng tích hợp và khả năng tương tác liền mạch giữa các giao thức khác nhau. Về tiêu chuẩn, một số tổ chức, như ITU-T, IETF, ISO / IEC, IEEE, ETSI, oneM2M và 3GPP, đã cho thấy một số nỗ lực. Mặc dù chúng tôi có thể giả định rằng tất cả các hoạt động tiêu chuẩn hóa này sẽ cung cấp khả năng tương tác (hoặc một số hình thức của nó), chúng có thể dẫn đến sự mơ hồ cao hơn vì thay vì có phạm vi rộng, tất cả đều cung cấp các giải pháp cụ thể và riêng biệt chỉ bao gồm các miền riêng.

Về nền tảng IoT, một số sáng kiến ​​đã được đưa ra để tạo ra lợi nhuận từ IoT bằng cách cung cấp kết nối, lưu trữ dữ liệu, phân tích dữ liệu lớn, dự đoán và học máy. Các công ty lớn trong ngành đã đạt được nhiều tiến bộ trong việc cung cấp các nền tảng IoT đa dạng với bộ tính năng phong phú. IBM Watson, Microsoft Azure, GE Predix, Cisco Jasper và PTC ThingWorx là những ví dụ về các nền tảng cấp doanh nghiệp phải đối mặt với việc khóa nhà cung cấp. Các sáng kiến ​​nền tảng IoT nguồn mở, chẳng hạn như Thingsboard.io, Kaa và DeviceHive, là một vài ví dụ điển hình trong vấn đề này.

1. **Thách thức về bảo mật**

Một trở ngại lớn trong việc tích hợp IoT rộng rãi trong các ngành công nghiệp nằm ở tính bảo mật của nó. Một số thách thức, chẳng hạn như cuộc tấn công Dyn gần đây, nhấn mạnh tầm quan trọng của việc có các thiết bị, nền tảng và ứng dụng IoT an toàn có thể dẫn đến những thảm họa lớn, chẳng hạn như thực hiện thành công một cuộc tấn công DDOS lớn. Những cuộc tấn công này có thể có tác động tàn phá đối với các doanh nghiệp của nhiều ngành công nghiệp quan trọng, đe dọa an ninh quốc gia và thậm chí ảnh hưởng trực tiếp hoặc gián tiếp đến cuộc sống của con người. Các chuyên gia CNTT trong các ngành này nắm trong tay các vấn đề bảo mật của BYOD và việc triển khai cơ sở hạ tầng đám mây tại chỗ trong các tổ chức của họ. Do đó, vấn đề bảo mật IoT chỉ làm tăng thêm sự lo lắng của họ. Bảo mật cũng không phải là chủ đề đầu tiên trong các cuộc thảo luận về IoT hiện tại và phần lớn vẫn được coi là một chủ đề bắt buộc nhưng thứ yếu. Sự coi thường như vậy có thể được quy cho việc thiếu các chính sách tổ chức và sự mơ hồ trong luật pháp của chính phủ. Để đảm bảo triển khai thành công IoT, việc giải quyết các vấn đề bảo mật này phải được ưu tiên trong lĩnh vực IoT. Những vấn đề này không chỉ đòi hỏi các giải pháp kỹ thuật mà còn thực thi chính sách và hướng dẫn phù hợp. Quan điểm của tất cả các bên liên quan trong IoT cũng phải được xem xét.

1. **Vấn đề quản lý dữ liệu**

Dữ liệu IoT là tài sản có giá trị. Với sự gia tăng theo cấp số nhân của số lượng thiết bị, hệ thống và quy trình IoT, các phương pháp mới, như Data Lakes, đã xuất hiện để xử lý dữ liệu lớn. Data Lakes lưu trữ dữ liệu có cấu trúc và không cấu trúc mà không có bất kỳ khái niệm nào được hình thành trước về cách những dữ liệu này sẽ được sử dụng sau đó. Điều này

Cách tiếp cận không áp dụng ánh xạ lược đồ hoặc ngôn ngữ truy vấn và có thể lưu trữ bất kỳ dữ liệu nào mà không bị hạn chế. Tuy nhiên, Data Lakes giới thiệu một vài vấn đề. Đầu tiên, do có thể chèn bất kỳ dữ liệu nào, việc hoán đổi dữ liệu có thể xảy ra trong tương lai. Để tránh vấn đề như vậy, chúng ta phải có sự giám sát về chất lượng dữ liệu, áp dụng bao gồm siêu dữ liệu và đảm bảo xuất xứ dữ liệu. Thứ hai, sử dụng Data Lakes có thể dẫn đến mất sự nhanh nhẹn, điều này đặc biệt đúng đối với các tổ chức lớn có ý định sử dụng một nhóm dữ liệu lớn để phân tích nhanh và đưa ra quyết định nhưng không thể thực hiện một cách hiệu quả vì họ phải trải qua vài bước trước đó trích xuất một cái gì đó có ý nghĩa từ dữ liệu. Thay vào đó, các tổ chức này phải phân biệt rõ ràng giữa những dữ liệu có thể được sử dụng để ra quyết định trong thời gian gần và những dữ liệu có thể được sử dụng để rút ra các chiến lược kinh doanh. Kiểu dữ liệu sau phù hợp hơn để lưu trữ trong Data Lake vì những dữ liệu này sẽ không được sử dụng ngay lập tức.

1. **Chứng minh dữ liệu**

Nguồn gốc dữ liệu được liên kết với tính xác thực và tính toàn vẹn của dữ liệu cũng như khả năng truy nguyên của chúng để xác định chủ sở hữu và người sửa đổi dữ liệu ở mỗi bước. Tuy nhiên, do dữ liệu lớn cung cấp những hiểu biết và phân tích sâu sắc có thể dẫn đến một số dạng hành động tự trị trong thế giới thực, chúng tôi phải đảm bảo rằng dữ liệu được sử dụng để thực hiện hành động đó là từ một nguồn hợp pháp. Một số sáng kiến ​​quy mô lớn, bao gồm các thành phố thông minh và sức khỏe thông minh, có kế hoạch sử dụng dữ liệu lớn và phân tích, do đó làm cho vấn đề này trở nên nghiêm trọng hơn. Mặc dù phần lớn các nghiên cứu hiện tại về IoT đã tập trung vào quản lý dữ liệu, nhưng chỉ có một số ít cố gắng giải quyết vấn đề xuất xứ dữ liệu, chẳng hạn như. Có khả năng theo dõi quyền sở hữu dữ liệu trong IoT có thể có lợi cho mục đích kiếm tiền khi các tác nhân khác nhau chia sẻ dữ liệu của họ. Các nghiên cứu hiện có trong lĩnh vực IoT, chẳng hạn như, có thể được sử dụng làm cơ sở để đưa ra các giải pháp kỹ thuật cho vấn đề này.

1. **Quản trị và điều chỉnh dữ liệu**

Một khía cạnh quan trọng của dữ liệu IoT có liên quan đến quản trị dữ liệu và điều chỉnh việc sử dụng nó bởi các thực thể khác nhau. Cung cấp quyền truy cập không được giám sát hoặc không được kiểm soát vào dữ liệu gây ra mối lo ngại về quyền riêng tư và cản trở sự tham gia của chủ sở hữu tư nhân, chẳng hạn như công dân chia sẻ dữ liệu của họ từ các cảm biến được cài đặt trong nhà hoặc ở nơi công cộng cho mục đích giám sát [81]. Chúng tôi phải cung cấp cho chủ sở hữu thiết bị IoT các tùy chọn và công cụ để chỉ định tùy chọn của họ và ưu tiên / giới hạn việc sử dụng dữ liệu từ thiết bị của họ [82] - [84]. Các nghiên cứu trong tương lai cũng phải tập trung vào việc phát triển các khung chính sách để xác định các cổ phần và mối quan tâm của chủ sở hữu dữ liệu, người tiêu dùng dữ liệu và tất cả các tác nhân khác giữa hai bên này. Đầu vào từ các cơ quan quản lý hoặc chính phủ sẽ là cần thiết, nhưng phải cẩn thận để không kiểm soát tập trung dữ liệu. Chủ sở hữu dữ liệu phải được trao thêm quyền lực để cho phép họ đưa ra quyết định trong phạm vi khung chính sách tổng thể. Công chúng phải được nhận thức về vai trò của họ và phải được cung cấp các công cụ dễ sử dụng để chia sẻ dữ liệu của họ với các bên khác.

1. **KẾT LUẬN**

IoT là một trong những nguồn dữ liệu lớn nhất, được kết xuất vô dụng nếu không có sức mạnh phân tích. IoT tương tác với dữ liệu lớn khi cần một lượng dữ liệu khổng lồ để xử lý, biến đổi và phân tích với tần suất cao. Công việc này đặc biệt tập trung vào bối cảnh dữ liệu lớn. Đầu tiên, chúng tôi điều tra các tài liệu gần đây về các giải pháp xử lý và phân tích dữ liệu lớn cho IoT. Thứ hai, chúng tôi xác định nhiều yêu cầu đối với dữ liệu lớn và phân tích trong IoT. Thứ ba, chúng tôi phân loại tài liệu. Thứ tư, chúng tôi xác định các cơ hội khác nhau được mang lại bởi dữ liệu lớn. Thứ năm, chúng tôi nhấn mạnh vai trò của phân tích dữ liệu trong các ứng dụng IoT. Thứ sáu, chúng tôi trình bày những thách thức nghiên cứu mở phải được giải quyết trong tương lai. Thứ bảy, chúng tôi kết luận rằng các giải pháp dữ liệu lớn hiện có trong mô hình IoT vẫn còn ở giai đoạn sơ khai và những thách thức liên quan đến chúng phải được giải quyết trong tương lai.

**LỜI CẢM ƠN**

Công việc này được hỗ trợ bởi Trưởng khoa nghiên cứu khoa học tại Đại học King Saud thông qua Nhóm nghiên cứu Imran từ số (RG # 1435-051)

**CHÚ THÍCH**

[1] I. Yaqoob, I. A. T. Hashem, A. Gani, S. Mokhtar, E. Ahmed, N. B. Anuar, and A. V. Vasilakos, “Big data: From begin- ning to future,” *International Journal of Information Manage- ment*, vol. 36, no. 6, pp. 1231–1247, 2016. [2] F. J. Riggins and S. F. Wamba, “Research directions on the adoption, usage, and impact of the internet of things through the use of big data analytics,” in *Proceedings of 48th Hawaii International Conference on System Sciences (HICSS’15)*. IEEE, 2015, pp. 1531–1540. [3] M. R. Bashir and A. Q. Gill, “Towards an iot big data an- alytics framework: Smart buildings systems,” in *High Per- formance Computing and Communications; IEEE 14th Interna- tional Conference on Smart City; IEEE 2nd International Con- ference on Data Science and Systems (HPCC/SmartCity/DSS), 2016 IEEE 18th International Conference on*. IEEE, 2016, pp. 1325–1332. [4] C. Lee, C. Yeung, and M. Cheng, “Research on iot based cyber physical system for industrial big data analytics,” in *Industrial Engineering and Engineering Management (IEEM), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1855–1859. [5] P. Rizwan, K. Suresh, and M. R. Babu, “Real-time smart traffic management system for smart cities by using in- ternet of things and big data,” in *Emerging Technological Trends (ICETT), International Conference on*. IEEE, 2016, pp. 1–7. [6] Q. Zhang, X. Zhang, Q. Zhang, W. Shi, and H. Zhong, “Firework: Big data sharing and processing in collabo- rative edge environment,” in *Hot Topics in Web Systems and Technologies (HotWeb), 2016 Fourth IEEE Workshop on*. IEEE, 2016, pp. 20–25. [7] M. M. Rathore, A. Ahmad, and A. Paul, “Iot-based smart city development using big data analytical approach,” in *Automatica (ICA-ACCA), IEEE International Conference on*. IEEE, 2016, pp. 1–8. [8] B. Ahlgren, M. Hidell, and E. C.-H. Ngai, “Internet of things for smart cities: Interoperability and open data,” *IEEE Internet Computing*, vol. 20, no. 6, pp. 52–56, 2016. [9] O. B. Sezer, E. Dogdu, M. Ozbayoglu, and A. Onal, “An extended iot framework with semantics, big data, and analytics,” in *Big Data (Big Data), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1849–1856. [10] B. Cheng, A. Papageorgiou, F. Cirillo, and E. Kovacs, “Geelytics: Geo-distributed edge analytics for large scale iot systems based on dynamic topology,” in *Internet of Things (WF-IoT), 2015 IEEE 2nd World Forum on*. IEEE, 2015, pp. 565–570. [11] H. Wang, O. L. Osen, G. Li, W. Li, H.-N. Dai, and W. Zeng, “Big data and industrial internet of things for the maritime industry in northwestern norway,” in *TENCON 2015-2015 IEEE Region 10 Conference*. IEEE, 2015, pp. 1–5. [12] J. L. P ́erez and D. Carrera, “Performance characterization of the servioticy api: an iot-as-a-service data management platform,” in *Big Data Computing Service and Applications (BigDataService), 2015 IEEE First International Conference on*. IEEE, 2015, pp. 62–71.

[13] M. Villari, A. Celesti, M. Fazio, and A. Puliafito, “Alljoyn lambda: An architecture for the management of smart environments in iot,” in *Smart Computing Workshops (SMARTCOMP Workshops), 2014 International Conference on*. IEEE, 2014, pp. 9–14. [14] A. J. Jara, D. Genoud, and Y. Bocchi, “Big data for cyber physical systems: an analysis of challenges, solutions and opportunities,” in *Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), 2014 Eighth International Conference on*. IEEE, 2014, pp. 376–380. [15] Z. Ding, X. Gao, J. Xu, and H. Wu, “Iot-statisticdb: a general statistical database cluster mechanism for big data analysis in the internet of things,” in *Green Computing and Communications (GreenCom), 2013 IEEE and Internet of Things (iThings/CPSCom), IEEE International Conference on and IEEE Cyber, Physical and Social Computing*. IEEE, 2013, pp. 535–543. [16] C. Vuppalapati, A. Ilapakurti, and S. Kedari, “The role of big data in creating sense ehr, an integrated approach to create next generation mobile sensor and wearable data driven electronic health record (ehr),” in *Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on*. IEEE, 2016, pp. 293–296. [17] A. Ahmad, M. M. Rathore, A. Paul, and S. Rho, “Defin- ing human behaviors using big data analytics in social internet of things,” in *Advanced Information Networking and Applications (AINA), 2016 IEEE 30th International Conference on*. IEEE, 2016, pp. 1101–1107. [18] E. Ahmed and M. H. Rehmani, “Introduction to the spe- cial section on social collaborative internet of things,” p. 382384, 2017. [19] D. Arora, K. F. Li, and A. Loffler, “Big data analytics for classification of network enabled devices,” in *Advanced In- formation Networking and Applications Workshops (WAINA), 2016 30th International Conference on*. IEEE, 2016, pp. 708– 713. [20] I.-L. Yen, G. Zhou, W. Zhu, F. Bastani, and S.-Y. Hwang, “A smart physical world based on service technologies, big data, and game-based crowd sourcing,” in *Web Services (ICWS), 2015 IEEE International Conference on*. IEEE, 2015, pp. 765–772. [21] R. P. Minch, “Location privacy in the era of the internet of things and big data analytics,” in *System Sciences (HICSS), 2015 48th Hawaii International Conference on*. IEEE, 2015, pp. 1521–1530. [22] A. Mukherjee, H. S. Paul, S. Dey, and A. Banerjee, “Angels for distributed analytics in iot,” in *Internet of Things (WF- IoT), 2014 IEEE World Forum On*. IEEE, 2014, pp. 565–570. [23] A. Mukherjee, S. Dey, H. S. Paul, and B. Das, “Utilis- ing condor for data parallel analytics in an iot contex- tan experience report,” in *Wireless and Mobile Computing, Networking and Communications (WiMob), 2013 IEEE 9th International Conference on*. IEEE, 2013, pp. 325–331. [24] H. R. Arkian, A. Diyanat, and A. Pourkhalili, “Mist: Fog- based data analytics scheme with cost-efficient resource provisioning for iot crowdsensing applications,” *Journal of Network and Computer Applications*, vol. 82, pp. 152–165, 2017. [25] M. M. Rathore, A. Ahmad, A. Paul, and S. Rho, “Urban planning and building smart cities based on the internet of things using big data analytics,” *Computer Networks*, vol. 101, pp. 63–80, 2016. [26] F. Alam, R. Mehmood, I. Katib, and A. Albeshri, “Analysis of eight data mining algorithms for smarter internet of

things (iot),” *Procedia Computer Science*, vol. 98, pp. 437– 442, 2016. [27] M. H. Berlian, T. E. R. Sahputra, B. J. W. Ardi, L. W. Dzatmika, A. R. A. Besari, R. W. Sudibyo, and S. Sukarid- hoto, “Design and implementation of smart environment monitoring and analytics in real-time system framework based on internet of underwater things and big data,” in *Electronics Symposium (IES), 2016 International*. IEEE, 2016, pp. 403–408. [28] D. Mourtzis, E. Vlachou, and N. Milas, “Industrial big data as a result of iot adoption in manufacturing,” *Procedia CIRP*, vol. 55, pp. 290–295, 2016. [29] R. Ramakrishnan and L. Gaur, “Smart electricity distri- bution in residential areas: Internet of things (iot) based advanced metering infrastructure and cloud analytics,” in *Internet of Things and Applications (IOTA), International Conference on*. IEEE, 2016, pp. 46–51. [30] E. Ahmed and M. H. Rehmani, “Mobile edge computing: Opportunities, solutions, and challenges,” pp. 59–63. [31] A. Ahmed and E. Ahmed, “A survey on mobile edge computing,” in *Intelligent Systems and Control (ISCO), 2016 10th International Conference on*. IEEE, 2016, pp. 1–8. [32] U. Shaukat, E. Ahmed, Z. Anwar, and F. Xia, “Cloudlet deployment in local wireless networks: Motivation, ar- chitectures, applications, and open challenges,” *Journal of Network and Computer Applications*, vol. 62, pp. 18–40, 2016. [33] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, “Fog com- puting and its role in the internet of things,” in *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*. ACM, 2012, pp. 13–16. [34] J. Nandimath, E. Banerjee, A. Patil, P. Kakade, S. Vaidya, and D. Chaturvedi, “Big data analysis using apache hadoop,” in *Information Reuse and Integration (IRI), 2013 IEEE 14th International Conference on*. IEEE, 2013, pp. 700– 703. [35] I. A. T. Hashem, N. B. Anuar, A. Gani, I. Yaqoob, F. Xia, and S. U. Khan, “Mapreduce: Review and open chal- lenges,” *Scientometrics*, pp. 1–34, 2016. [36] V. Morabito, “Managing change for big data driven inno- vation,” in *Big Data and Analytics*. Springer, 2015, pp. 125–153. [37] A. Bhardwaj, S. Bhattacherjee, A. Chavan, A. Deshpande, A. J. Elmore, S. Madden, and A. G. Parameswaran, “Datahub: Collaborative data science & dataset version management at scale,” *arXiv preprint arXiv:1409.0798*, 2014. [38] F. F ̈arber, S. K. Cha, J. Primsch, C. Bornh ̈ovd, S. Sigg, and W. Lehner, “Sap hana database: data management for modern business applications,” *ACM Sigmod Record*, vol. 40, no. 4, pp. 45–51, 2012. [39] S. Burke, “Hp haven big data platform is gain- ing partner momentum,” *CRN [online] http://www. crn. com/news/applications-os/240161649*, 2013. [40] (2017, Accessed on 3rd June) Hortonworks. [Online].

Available: https://hortonworks.com/ [41] Y. Zhuang, Y. Wang, J. Shao, L. Chen, W. Lu, J. Sun, B. Wei, and J. Wu, “D-ocean: an unstructured data management system for data ocean environment,” *Frontiers of Computer Science*, vol. 10, no. 2, pp. 353–369, 2016. [Online]. Available: http://dx.doi.org/10.1007/s11704-015-5045-6 [42] D. Slezak, P. Synak, J. Wr ́oblewski, and G. Toppin, “In- fobright analytic database engine using rough sets and granular computing,” in *Granular Computing (GrC), 2010 IEEE International Conference on*. IEEE, 2010, pp. 432–437. [43] (2017, Accessed on 3rd June) Mapr. [Online]. Available:

https://mapr.com/

[44] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al- Jaroodi, “Applications of big data to smart cities,” *Journal of Internet Services and Applications*, vol. 6, no. 1, p. 1, 2015. [45] E. Ahmed, M. Imran, M. Guizani, A. Rayes, J. Lloret, G. Han, and W. Guibene, “Enabling mobile and wireless technologies for smart cities: Part 2,” *IEEE Communications Magazine*, vol. 55, no. 3, pp. 12–13, 2017. [46] G. Suciu, V. Suciu, A. Martian, R. Craciunescu, A. Vulpe, I. Marcu, S. Halunga, and O. Fratu, “Big data, internet of things and cloud convergence–an architecture for secure e-health applications,” *Journal of medical systems*, vol. 39, no. 11, pp. 1–8, 2015. [47] J. Jin, J. Gubbi, T. Luo, and M. Palaniswami, “Network architecture and qos issues in the internet of things for a smart city,” in *Communications and Information Technologies (ISCIT), 2012 International Symposium on*. IEEE, 2012, pp. 956–961. [48] R. T ̈onjes, P. Barnaghi, M. Ali, A. Mileo, M. Hauswirth, F. Ganz, S. Ganea, B. Kjærgaard, D. Kuemper, S. Nechifor *et al.*, “Real time iot stream processing and large-scale data analytics for smart city applications,” in *poster session, European Conference on Networks and Communications*, 2014. [49] E. Ahmed, S. Ali, A. Akheenzada, and I. Yaqoob, “Cogni- tive radio sensor networks: Bridging the gap for network,” *Cognitive Radio Sensor Networks: Applications, Architectures, and Challenges: Applications, Architectures, and Challenges. IGI Global*, p. 160, 2014. [50] S. A. A. Shah, E. Ahmed, F. Xia, A. Karim, M. A. Qureshi, I. Ali, and R. M. Noor, “Coverage differentiation based adaptive tx-power for congestion and awareness control in vanets,” *Mobile Networks and Applications*, pp. 1–12. [51] I. Yaqoob, I. Ahmad, E. Ahmed, A. Gani, M. Imran, and N. Guizani, “Overcoming the key challenges to establish- ing vehicular communication: Is sdn the answer?” *IEEE Communications Magazine*, 2017. [52] I. Yaqoob, E. Ahmed, I. A. T. Hashem, A. Ahmed, A. Gani, M. Imran, and M. Guizani, “Internet of things architec- ture: Recent advances, taxonomy, requirements, and open challenges,” *IEEE Wireless Communications*, 2017. [53] Y. Jararweh, A. Doulat, O. AlQudah, E. Ahmed, M. Al- Ayyoub, and E. Benkhelifa, “The future of mobile cloud computing: integrating cloudlets and mobile edge com- puting,” in *Telecommunications (ICT), 2016 23rd Interna- tional Conference on*. IEEE, 2016, pp. 1–5. [54] N. Bessis and C. Dobre, *Big data and internet of things: a*

*roadmap for smart environments*. Springer, 2014. [55] I. A. T. Hashem, V. Chang, N. B. Anuar, K. Adewole, I. Yaqoob, A. Gani, E. Ahmed, and H. Chiroma, “The role of big data in smart city,” *International Journal of Information Management*, vol. 36, no. 5, pp. 748–758, 2016. [56] A. Brring, S. Schmid, C. K. Schindhelm, A. Khelil, S. Kbisch, D. Kramer, D. L. Phuoc, J. Mitic, D. Anicic, and E. Teniente, “Enabling iot ecosystems through platform interoperability,” *IEEE Software*, vol. 34, no. 1, pp. 54–61, Jan 2017. [57] X. W. Chen and X. Lin, “Big data deep learning: Chal- lenges and perspectives,” *IEEE Access*, vol. 2, pp. 514–525, 2014. [58] J. Qiu, Q. Wu, G. Ding, Y. Xu, and S. Feng, “A survey of machine learning for big data processing,” *EURASIP Journal on Advances in Signal Processing*, vol. 2016, no. 1, pp. 1–16, 2016. [59] O. Etzion, “When artificial intelligence meets the internet of things,” in *Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems*. ACM, 2015, pp. 246–246.

[60] V. O. Safonov, “Example of a trustworthy cloud comput- ing platform in detail: Microsoft azure,” *Trustworthy Cloud Computing*, pp. 147–270, 2016. [61] J. Vidal-Garcıa, M. Vidal, and R. H. Barros, “Computa- tional business intelligence, big data, and their role in business decisions in the age of the internet of things,” in *The Internet of Things in the Modern Business Environment*. IGI Global, 2017, pp. 249–268. [62] Y. Jeong, H. Joo, G. Hong, D. Shin, and S. Lee, “Aviot: Web-based interactive authoring and visualization of in- door internet of things,” *IEEE Transactions on Consumer Electronics*, vol. 61, no. 3, pp. 295–301, 2015. [63] M. Strohbach, H. Ziekow, V. Gazis, and N. Akiva, “To- wards a big data analytics framework for iot and smart city applications,” in *Modeling and processing for next- generation big-data technologies*. Springer, 2015, pp. 257– 282. [64] S. Aljawarneh, V. Radhakrishna, P. V. Kumar, and V. Janaki, “A similarity measure for temporal pattern dis- covery in time series data generated by iot,” in *Engineering & MIS (ICEMIS), International Conference on*. IEEE, 2016, pp. 1–4. [65] C. El Kaed, I. Khan, H. Hossayni, and P. Nappey, “Sqeniot: Semantic query engine for industrial internet-of-things gateways,” *Submitted IEEE GLOBECOM*, 2016. [66] T. Banerjee and A. Sheth, “Iot quality control for data and application needs,” *IEEE Intelligent Systems*, vol. 32, no. 2, pp. 68–73, 2017. [67] A. A. Bhattacharya, D. Hong, D. Culler, J. Ortiz, K. White- house, and E. Wu, “Automated metadata construction to support portable building applications,” in *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. ACM, 2015, pp. 3–12. [68] T. Jaffey, J. Davies, and P. Beart, “Hypercat 3.00 specifica-

tion,” *Hyper-cat Limited*, 2016. [69] P. Desai, A. Sheth, and P. Anantharam, “Semantic gateway as a service architecture for iot interoperability,” in *Mobile Services (MS), 2015 IEEE International Conference on*. IEEE, 2015, pp. 313–319. [70] A. Meddeb, “Internet of things standards: who stands out from the crowd?” *IEEE Communications Magazine*, vol. 54, no. 7, pp. 40–47, 2016. [71] C. Chris Mishler and C. CIA, “The future of the internet of things,” *Strategic Finance*, vol. 97, no. 5, p. 62, 2015. [72] S. Tanimoto, S. Yamada, M. Iwashita, T. Kobayashi, H. Sato, and A. Kanai, “Risk assessment of byod: Bring your own device,” in *Consumer Electronics, 2016 IEEE 5th Global Conference on*. IEEE, 2016, pp. 1–4. [73] K. Hajdarevic, P. Allen, and M. Spremic, “Proactive se- curity metrics for bring your own device (byod) in iso 27001 supported environments,” in *Telecommunications Fo- rum (TELFOR), 2016 24th*. IEEE, 2016, pp. 1–4. [74] V. A. Almeida, D. Doneda, and J. de Souza Abreu, “Cyber- warfare and digital governance,” *IEEE Internet Computing*, vol. 21, no. 2, pp. 68–71, 2017. [75] H. Fang, “Managing data lakes in big data era: What’s a data lake and why has it became popular in data management ecosystem,” in *Cyber Technology in Automa- tion, Control, and Intelligent Systems (CYBER), 2015 IEEE International Conference on*. IEEE, 2015, pp. 820–824. [76] R. Hai, S. Geisler, and C. Quix, “Constance: An intelligent data lake system,” in *Proceedings of the 2016 International Conference on Management of Data*. ACM, 2016, pp. 2097– 2100.

19

[77] W. She, I.-L. Yen, F. Bastani, B. Tran, and B. Thurais- ingham, “Role-based integrated access control and data provenance for soa based net-centric systems,” in *Service Oriented System Engineering (SOSE), 2011 IEEE 6th Interna- tional Symposium on*. IEEE, 2011, pp. 225–234. [78] B. Glavic, “Big data provenance: Challenges and implica- tions for benchmarking,” in *Specifying big data benchmarks*. Springer, 2014, pp. 72–80. [79] Y. Zhang and J. Wen, “An iot electric business model based on the protocol of bitcoin,” in *Intelligence in Next Genera- tion Networks (ICIN), 2015 18th International Conference on*. IEEE, 2015, pp. 184–191. [80] Q. H. Cao, I. Khan, R. Farahbakhsh, G. Madhusudan, G. M. Lee, and N. Crespi, “A trust model for data shar- ing in smart cities,” in *Communications (ICC), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–7. [81] I. Khan, F. Belqasmi, R. Glitho, N. Crespi, M. Morrow, and P. Polakos, “Wireless sensor network virtualization: Early architecture and research perspectives,” *IEEE Network*, vol. 29, no. 3, pp. 104–112, 2015. [82] E. D. Ragan, A. Endert, J. Sanyal, and J. Chen, “Charac- terizing provenance in visualization and data analysis: an organizational framework of provenance types and purposes,” *IEEE transactions on visualization and computer graphics*, vol. 22, no. 1, pp. 31–40, 2016. [83] C. H. Suen, R. K. Ko, Y. S. Tan, P. Jagadpramana, and B. S. Lee, “S2logger: End-to-end data tracking mechanism for cloud data provenance,” in *Trust, Security and Privacy in Computing and Communications (TrustCom), 2013 12th IEEE International Conference on*. IEEE, 2013, pp. 594–602. [84] M. B. Jones, B. Lud ̈ascher, T. McPhillips, P. Missier, C. Schwalm, P. Slaughter, D. Vieglais, L. Walker, and Y. Wei, “Dataone: A data federation with provenance sup- port,” in *Provenance and Annotation of Data and Processes: 6th International Provenance and Annotation Workshop, IPAW 2016, McLean, VA, USA, June 7-8, 2016, Proceedings*, vol. 9672. Springer, 2016, p. 230.